**Title**
Accuracy of Professional Self-Reports: Medical Student Self-Report and the Scoring of Professional Competence



**Author**
Richter Lagha, Regina Anne

**Publication Date**
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# Accuracy of Professional Self-Reports:

# Medical Student Self-Report and the Scoring of Professional Competence

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Education

by

Regina Anne Richter Lagha

2014

ABSTRACT OF THE DISSERTATION

Accuracy of Professional Self-Reports:

Medical Student Self-Report and the Scoring of Professional Competence

by

Regina Anne Richter Lagha

Doctor of Philosophy in Education

University of California, Los Angeles, 2014

Professor Noreen M. Webb, Chair

Self-report is currently used as an indicator of professional practice in a variety of fields, including medicine and education. Important to consider, therefore, is the ability of self-report to accurately capture professional practice. This study investigated how well professionals' self-reports of behavior agreed with an expert observer's reports of those same behaviors. While this study explored self-report in the context of medical professionals, this topic is equally important to the measurement of teacher practices.

This study investigated agreement between: 1) medical student self-report and expert rater documentation of a clinical encounter; and 2) standardized patient (an actor highly-trained to portray a patient) and expert rater documentation of medical student performance. Additionally, this study investigated whether levels of agreement depended on the context

and content of behaviors, features of the examination, or characteristics of the professional.

Performance data were analyzed from a stratified random sample of 75 fourth-year medical students who completed a clinical competence examination in 2012. Students rotated through a series of 15-minute encounters, called stations, interviewing a standardized patient in each. Medical students were instructed to: 1) obtain the patient's history; 2) conduct a physical examination; and 3) discuss potential diagnoses. Ratings of student performance were collected from the medical student self-reports, the SP checklists, and the expert rater's documentation of the encounters. Analyses focused on the 4-7 behavioral items in each of the three stations studied that were considered critical to patient care.

Comparison of the three sources of ratings revealed marked differences. Most importantly, medical students' self-reports did not agree highly with the expert's reports. Medical students both under-reported and over-reported a substantial number of critical action items with level of agreement varying by station and nature of the behavior. Due to medical student tendency to under-report behaviors, use of self-report to score performance would result in a large number of students falsely identified as failing the examination.

This study discusses causes of medical student under- and over-report and recommends strategies for improvement. The study also addresses implications of findings for the use of self-report among teachers, citing specific examples.

The dissertation of Regina Anne Richter Lagha is approved.

Charles Goodwin

José Felipe Martínez

Mike A. Rose

Noreen M. Webb, Committee Chair

University of California, Los Angeles

2014

# DEDICATION

For the sake of Allah, my Creator and Master.

For my great teacher and messenger, Mohammed (peace be upon him),

who taught us the purpose of life.

For my mother, under whose feet heaven lies.

# Contents

# List of Figures

# List of Tables

# ACKNOWLEDGMENTS

In the Name of Allah, the Most Merciful, the Most Compassionate.

First and foremost, I give thanks to Allah, the Ever-Magnificent, All-Knowing, and Forgiving. This work would have been impossible without His guidance, help, and blessings.

I thank those who have mentored me over the years, particularly LuAnn Wilkerson and her incredible team at the David Geffen School of Medicine, Center for Education Development & Research, including Cha-Chi Fung, Elizabeth O'Gara, Sebastian Uijtdehaage, Paul Wimmers, Ming Lee Chung, Lawrence "Hy" Doyle, Arleen Brown, Art Gomez, and David Lazarus and as well as all those who offered friendship, support, advice, and laughs while working in the ED&R office.

I thank Adina Kalet, Linda Tewksbury, and Marian Anderson at NYU School of Medicine for taking a chance on a young college graduate.

I am also indebted to my friends and family and to my amazing (and very patient) husband–my rock, my companion, my reality check.

Finally, the research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B080016 to the University of California, Los Angeles. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

# Regina Anne Richter Lagha

## Professional Experience

**Education Researcher, David Geffen School of Medicine, Los Angeles, CA 10/07-6/12**

- Provided research support for the UCLA Hispanic Center of Excellence
- Conducted analyses for competency education and training research
- Assisted Clinical Education Task Force in its recommendations for curricular improvement

**Program Coordinator, NYU School of Medicine, New York, NY 9/05-8/07**

- Supported activities of Research on Medical Education Outcomes (ROMEO), which aimed to create evidence-base linking medical training to patient outcomes
- Conducted curriculum and evaluation methods needs assessment across residency programs
- Organized annual fourth-year medical student Comprehensive Clinical Skills Exam
- Provided grant writing, assisted with implementation of project evaluations, and data management
- Coauthored abstracts, posters, papers, and workshop presentations
- Facilitated the Primary Care and Public Health Scholars Program
- Assisted in grant writing process and act as liaison to NYU SoM Internal Review Board (IRB)

## Teaching Experience

**Special Reader, Ethnographic Field Methods, UCLA, Los Angeles, CA Winter 2009**

- Critiqued students' weekly field notes and offered guidance in the sense-making process
- Delivered portions of course lectures during quarter

## Education

**New York University, New York, NY** **June 2007**
*Masters in Museum Studies*

**Stanford University, Stanford, CA** **June 2003**
*Bachelor of Arts in Anthropological Sciences (with honors) and French Studies*

- Graduated Phi Beta Kappa with honors and distinction

## Selected publications

1. Richter Lagha, R. A., Boscardin, C. K., May, W., Fung, C. (2012). A comparison of two standard-setting approaches in high-stakes clinical performance assessment using generalizability theory. *Academic Medicine, 87*, 1077–1082.

2. Fung, C. C., Richter Lagha, R. A., Henderson, P., Gomez, A. G. (2010). Working with interpreters: How student behavior affects quality of patient interaction when using interpreters. *Medical Education Online, 15*, 5151. doi:10.3402/meo.v15i0.5151.

## Selected Peer Reviewed Abstracts, Papers, and Workshops

1. Richter Lagha, R. (2012, April). Estimating the reliability of performance scores: important considerations. Paper presented at the meeting of the American Educational Research Association, Vancouver, Canada.
2. Richter Lagha, R., Brown, A., Gomez, A. (2012, April). Modeling cultural competency: student-perceived qualities of culturally competent providers. Poster session presented at the meeting of the American Educational Research Association, Vancouver, Canada.
3. Richter Lagha, R., O'Gara E.M. (2011, November). Evaluating student oral presentation skills: reliability and validity of a standardized resident measure. Paper presented at the meeting of Research in Medical Education, Denver, CO.
4. O'Gara, E. M., Richter Lagha, R. (2011, November) Evaluating student oral presentation skills: development of a standardized resident. Poster session presented at the meeting of Research in Medical Education, Denver, CO.
5. Richter Lagha, R. (2011, April). The effect of clinical training for healthcare professionals on patient outcomes: a meta-analysis. Paper presented at the meeting of the American Education Research Association, New Orleans, LA.
   Medical Education, Los Angeles, CA.
   Education, Pasadena, CA.
6. Tewksbury, L., Richter R., Kalet, A. (2006, May). Remediating students with poor clinical skills performance: who are they and what can we do for them? Workshop session presented at the International Ottawa Conference on Clinical Competence, New York, NY.
7. Tewksbury, L., Gillespie, C., Richter, R., Kalet, A. (2006, April). Medical students with lowest performance on a clinical skills examination poorly self-assess ability. Poster session at the meeting of the Society of General Internal Medicine, Los Angeles, CA.

## Awards and Honors

- Advanced Qualitative Methods Fellowship, University of California, Los Angeles
- Chancellor's Prize, University of California, Los Angeles
- Excellence in Writing, Department of Anthropological Science, Stanford University
- Stanford University Women's Honor Society, Cap & Gown

## Skills and Interests

- Computer Skills: Macintosh/PC platforms; SPSS; HLM; LISREL; R; GENOVA
- Language Skills: Fluent in French; Conversational in Spanish; Beginning Arabic
- President, Graduate Student Association in Education, 2009-2010

# Chapter 1

# Introduction

Among both emerging and practicing professionals, such as teachers, professors, lawyers, social workers, and health professionals, there is a need for comprehensive, accurate assessment of performance to ensure the maintenance and improvement of lifelong skills. Self-documentation of practices and behaviors, also known as self-report, is an important practice in many professions, serving as a record of events and as a means of self-reflection. This study examines the accuracy of self-report data in order to contribute to our understanding of professional self-report, its strengths and weaknesses, and its use in the evaluation of professionals and the improvement of their performance. Study findings contribute to the development of best practices in the use of self-report among professionals, like educators, and contributes to a broader area of research, the measurement of teaching.

Self-report makes it possible to not only collect information about professional practices but also to gain insight into professional intent and interpretation of events not otherwise documented when using techniques like independent observation. This enables evaluators to identify areas in need of improvement not just in professional practice but in professional self-

perception of those practices. What is more, incorporating multiple methods of professional assessment of performance (e.g., self-report, peer-evaluation, supervisor ratings, etc.) into an assessment plan allows evaluators to correlate, even triangulate, performance measures, lending credibility to the assessment (or identifying its limitations). Finally, self-report can be less resource-consuming than other methods of gathering information about professional practice and behaviors both in terms of time and money, making it advantageous in comparison to other forms of assessment like trained observation. While self-report is an important component of professional evaluation and development, its accuracy is often questioned, and research specifically on the practice of professional self-assessment has done little to assuage concerns regarding its continued use. In examining the practice of self-report, this study investigates a particular methodology commonly used in the assessment of professionals, using data collected from a sample of medical professionals-in-training, to contribute to the development, implementation, and optimal use of self-report.

## 1.1 Self-report in education: an overview of teacher self-report

In education, the self-report can consist of teachers' documentation of their practices and behaviors in the classroom and self-evaluation of their own performance as well as documentation of student behaviors and interpretation of student performance. Teacher self-reports often figure into evaluation of teachers instructional practices alongside a variety of other methods, including: student ratings, student outcomes, observer ratings, peer ratings,

and supervisor ratings (Centra, 1982; Doyle & Webber, 1978; Howard, Conway, & Maxwell, 1985; Kulik & McKeachie, 1975; O'Hanlon & Mortensen, 1980). Centra (1982) reported a growing trend in the use of self-report as part of college and university tenure evaluations as well as for faculty development and teacher improvement. Previous research, however, has raised questions about the accuracy of teachers' self-reports compared to assessments made by students, outside observers, peers, and supervisors (Blackburn & Clark, 1975; Doyle & Crichton, 1978; Howard et al., 1985; Marsh, Overall, & Kesler, 1979; Webb & Nolan, 1955).

Arguments for self-report center around the simple fact that the teacher is in the best position to accurately document what he or she actually did and is more aware of the level of preparation and expertise he or she brought to the class (Kulik & McKeachie, 1975). As an example, the Carnegie-Mellon University English department, as detailed by Eastman (1970), gave faculty the option of self-report, student ratings, or classroom observation as part of the faculty review process. In the first year, few faculty chose the self-report option; however, for the ones who did, they provided rich, detailed, summative descriptions of their classroom practices and evaluation of their effectiveness. However, self-report is not without its limitations, mainly the concern of teacher misperception of teaching effectiveness (Blackburn & Clark, 1975; Centra, 1982).

Previous research is inconsistent about whether self-report data detailing teacher practices and behaviors agree with other sources of information, such as expert raters' observations and students' reports of teacher practices. Hartman and Nelson's study (1992) of medical school faculty found significant discrepancies between faculty self-report of instructional practices and performance as measured by written simulated teaching situations. In their study of teacher individualization of instruction, Hardebeck, Ashbaugh, and McIntyre

(1974) found that teacher self-report scores were significantly higher than those obtained from trained observers of teacher behavior using a detailed instrument. Measuring the validity of measures captured by a web-based teacher's log, Ball, Camburn, Correnti, Phelps, and Wallace (1999) found some discrepancy between teachers and observers in reporting the duration of the lesson, marking what the teacher was doing, and providing analysis of reading and math lessons. The authors felt further refinement of the log was needed to enhance agreement. Researchers with the Study of Instructional Improvement also found discrepancy between literacy teachers using a teacher log and researcher-observers, particularly regarding the documentation of certain focus areas, like reading comprehension and writing (Camburn & Barnes, 2004). Looking to improve the accuracy of self-reporting among high school teachers, Koziol and Burns (1986) found that the repeated use of a focused self-report tool improved teacher self-report accuracy based on comparison to reports by trained observers. Newfield (1980) also found in his study of elementary and middle school teachers significant correlation between teacher and observer report of teacher behaviors, noting that further research was needed to better understand what conditions might influence teacher accuracy.

It is important to distinguish self-report of practices from self-report of teaching effectiveness, or self-evaluation, which is a common practice though less pertinent to the present study. The self-evaluation literature also reveals inconsistencies, making it impossible to conclude definitively the ability of teachers to accurately self-evaluate their performance in comparison to students, outside observers, peers, even supervisors. In a study of teacher effectiveness of college instructors, Howard et al. (1985) used several different methods, including correlations of ratings between different rater groups and confirmatory factor analysis, to investigate the validity of common measures of teacher performance such as: teacher self-

evaluation, ratings by current students, ratings by former students, ratings by colleagues, and ratings by trained observers. The authors found higher validity coefficients (Pearson correlation coefficients, factor loadings) for current and former student ratings than self-evaluation, ratings by colleagues, and ratings by trained observers. The authors could not explain the low validity coefficients of teacher self-evaluation. In a study of faculty at a liberal arts college, Blackburn and Clark (1975) also found low correlation between faculty self-evaluation and student ratings. Some studies, however, have reported significant positive relationships between faculty self-evaluation and student ratings (Doyle & Crichton, 1978; Marsh et al., 1979; Webb & Nolan, 1955) These studies all demonstrate that self-evaluation of effectiveness, like self-report of practices, can be incorrect and that potential contributors to this inaccuracy are not well understood.

Teacher self-report, therefore, remains a topic of considerable interest. Several funded Institute of Education Sciences (IES) studies are currently either employing teacher self-report or investigating associated methodological issues to improve its usage. In a project titled "Scientific Validation of a Set of Instruments Measuring Fidelity of Implementation (FOI) of Reform-Based Science and Mathematics Instructional Materials" (Award No: R305A1100621), principal investigator Dae Kim of the University of Chicago investigates the reliability and validity of different tools used to measure how well the enacted curriculum (the implementation of the curriculum) adheres to the actual curriculum (to instructional materials as designed). One of the instruments under scrutiny is the instructional log, a particular form of teacher self-report. As with the present study, Kim is concerned with agreement between the different instruments and how contextual factors influence the psychometric properties of these instruments. Another IES funded project "Improving Teachers' Moni-

toring of Learning" (Award No: R305A120265) from principal investigator Keith Thiede at Boise State University posits that effective learning depends on teacher ability to monitor student learning accurately. In addition to observation and student achievement scores, the researchers plan on piloting the use of teacher self-report. If the practice is found to enhance the program in Stage 1, then it will be included in the final program. Principal investigator Maria Ruiz-Primo from the University of Colorado, Denver is also investigating the psychometric properties of teacher self-report in the measurement of formative assessment in the project "Developing and Evaluating Measures of Formative Assessment Practices" (Award No: R305A100571). She proposes to validate surrogate measures, such as self-report teacher log protocols, which are more cost-, time-, and effort-efficient than gold standard benchmark measures such as classroom observation protocols by a third-party. More information about self-report may lead to its usage in other projects, such as the IES funded National Research and Development Center on Scaling Up Effective Schools (Award No: R305C100023), which is attempting to identify practices of effective high schools with historically low-performing schools. Teacher self-report may serve an important role in such a project.

## 1.2 A parallel model of professional self-report: physician self-report

In the medical arena, self-reports are a common standard practice within the profession, comprising a crucial component of patients' permanent medical records. Physicians routinely employ self-report, in the form of the patient note, in order to record clinical encounters

with patients, communicate with other members of the medical team about a patient's care, and for billing purposes. Because physician self-report serves such a central role in the day to day practice of medicine, it provides a unique opportunity for the study of professional self-report. Medical schools commonly provide standardized mechanisms for the collection and analysis of these data. Though the practice of self-report is common within medicine and therefore a standard component of medical training, little is known about the accuracy of these self-reports; in particular, how well do they agree with other sources of information about professional behavior, such as expert raters' observations. This study, then, investigates important issues in self-report, namely: 1) what is the level of agreement between professional and expert rater in the documentation of a professional experience, in this instance, the report of a medical professional-in-training's performance in a clinical encounter; 2) how does that level of agreement between professional and expert compare to the level of agreement between a trained observer and the expert; and 3) do contextual factors matter to ability to correctly report performance in the professional self-report? In this study, routinely captured performance data, including professional self-report, trained observer and expert observation data, and video recording, were all used to establish professional ability to correctly self-report performance. Is ability to properly self-report tied to features of the professional? Or to aspects of the experience, in this case the clinical encounter, but in the case of teachers, the classroom encounter?

The results of this study will have direct implications for the use of professional self-report in education in two areas: 1) teacher evaluation, where self-report can be one of several different methods used to assess performance, teacher competence, and enactment of new curriculum; and 2) instructional improvement, where self-report can be used to document

teaching practices and teacher cognition to aid in enhancing teacher efficacy. Chapter 6 will address these implications in detail.

## 1.3 Statement of problem: assessing medical student self-report ability

Medical educators task themselves with producing clinically competent physicians, armed with the knowledge, skills and values necessary to serve not just individuals, but the entire community, entrusted to their care. To ensure competence in their graduates, these same medical institutions have come to rely on performance-based assessment to determine medical student proficiency in a complex construct–clinical competence–that lies at the heart of patient health, safety, and care. This study addressed a primary concern of performance-based assessment in medical education: score accuracy.

Developing reliable and valid assessment measures of clinical competence is a key concern of medical institutions. Many clinical competence assessments are marred by score reliability and validity concerns (de Champlain, Margolis, King, & Klass, 1997; Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2007; Newble & Swanson, 1988; Stilson, 2009; Tamblyn, 1989; van der Vleuten & Swanson, 1990), and only a very few, select studies have demonstrated a relationship between these measures and actual practice (Hamdy et al., 2006; Pieters, Touw-Otten, & de Melker, 2002; Tamblyn et al., 2002). As institutions adopt new methods of gathering information about medical student performance, they also introduce potential new sources of error in the interpretation of medical student behaviors and clinical

skills. How best to determine clinical competence, or improve on existing measures, therefore endures as an important topic of study.

Designed as a more rigorous, criteria-based, and standardized measure of clinical performance, the Objective Structured Clinical Examination (OSCE) involves the rotation of medical students through a series of small, focused clinical cases, or stations. In each station, medical students interact with a standardized patient (SP), or an actor trained to simulate a specific illness in a consistent fashion. Evaluation of medical student performance is based on information provided by a rater. This can include a faculty observer, trained non-faculty expert, SP trainer, peer, the SP in the encounter, or the medical student him/herself. Though ratings of medical student behavior provided by SPs oftentimes form the core of clinical competence assessment, use of information about the encounter provided by the medical student is gaining popularity. Although the accuracy of SP ratings of the medical student's behavior during such encounters has been studied, little is known about medical student ability to report performance in an encounter. This study, then, examines the veracity of information provided by the medical student in one such examination.

Generally during an OSCE, SPs use checklists to note the presence or absence of a desired behavior in a simulated clinical encounter. Though ample evidence does exist demonstrating SP ability to furnish accurate scores (de Champlain et al., 1997; Henry & Smith, 2010; Pangaro, Worth-Dickstein, MacMillan, Klass, & Shatzer, 1997; van Zanten, Boulet, McKinley, & Whelan, 2003; Williams, 2004), some studies have found limitations to SP ability to correctly document student performance, for instance, in their inability to correctly rate medical student behaviors given a lengthy checklist or when using complex checklist items (Vu et al., 1992). Other studies have shown that SPs have difficulty correctly rating medical

students in certain aspects of the clinical encounter, like physical examination or patient education (de Champlain et al., 1997; Tamblyn, Grad, Gayton, Petrella, & Reid, 1997).

Few researchers have studied medical student ability to correctly document the student-SP encounter (Tamblyn et al., 1997). Following a clinical encounter, medical students complete a self-report, termed a "patient note," which documents their findings including: a) the patient's history based on the medical interview; b) a record of the physical examination as performed by the medical student and his or her findings; c) a differential diagnosis, or a list of potential diagnoses listed in order of most to least likely as justified by the information provided in the history and physical examination sections of the note; and d) a patient work-up, or list of next steps, including tests to be performed, to fully diagnose the patient.

Some large-scale testing agencies, however, like the National Board of Medical Examiners (NBME), have begun to use medical student documentation as a basis for scoring components of the student-SP encounter. As part of the national licensure examination, the United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills (Step 2 CS), students rotate through a series of twelve encounters with SPs and are required immediately following each encounter to document information obtained, physical examination findings, and information about potential diagnoses and next steps in diagnosis and treatment. Evaluators then use this information, documented by the medical student, as evidence of performed behaviors, such as asking important questions of the patient during the medical interview and examining the patient. Specifically, the "Integrated Clinical Encounter" component of the examination, which assesses student skills in data gathering and interpretation, relies on: 1) SP checklists of the physical examination performed in the encounter; and 2) student description of the history and physical examination, potential diagnoses, justification for those diagnoses, and

desired diagnostic studies in the patient note as rated by trained physicians (United States Medical Licensing Examination, 2013). As institutions begin to introduce this new method of scoring the clinical encounter and rely increasingly on medical student report of what did and did not happen in the encounter, they must address issues concerning medical student ability to properly recall and report the encounter. Ensuring veracity of performance data is of utmost importance, as scores derived from incorrect ratings can negatively impact score interpretation.

This study examines the veracity of information collected from medical students, relying on data obtained from a high-stakes OSCE at a local medical institution. Both medical student and SP documentation of examinee performance in the encounter were compared to expert ratings of medical student behavior. Also of concern to this study were contextual factors–like the content and context of information about behaviors collected, characteristics of the examination itself, and features of the professional– and their impact on medical student ability to recall performance correctly. This study examined the relationship between what medical students truly performed in the encounter and what they subsequently reported about the patient. By investigating potential contextual factors and the role they played in the ability of medical students to appropriately document the encounter, this study aims to better inform construction of assessments and interpretation of performance scores using self-report.

## 1.4 Purpose of performance-based assessment of clinical competence in medicine

In a 1990 invited lecture on the assessment of clinical skills, the late George Miller, one of the early pioneers in the field of medical education, started with a general disclaimer: "...it seems important to start with the forthright acknowledgement that no single assessment method can provide all the data required for judgment of anything so complex as the delivery of professional services by a successful physician" (G. E. Miller, 1990, p. S63). Miller outlined a four-tiered pyramid of clinical competence assessment comprised of "knows" or knowledge at the base, followed by "knows how" or competence, "shows how" or performance, and culminating in "does" or action.

According to Miller, at the most basic levels, medical educators should not only assess for medical student knowledge but for student ability to apply that knowledge. Medical students must show how well they use that "knows how" through assessment, and educators should concern themselves with how well medical student performance in artificial clinical examination environments predicts future routine clinical practice. With each subsequent level of Miller's assessment pyramid, uncertainty surrounding the reliability and validity of assessment measures increases, as methods transition from generally well-established "objective test methods," such as knowledge-based national licensing exams, which may capture medical student ability reflected in the lower levels of Miller's framework, to evaluations of patient interviews and simulated clinical encounters, necessary to capture medical student ability in the upper levels of Miller's framework. To this day, medical educators continue to investigate the value of clinical skills assessment as an accurate and consistent measure of

medical student clinical performance.

Clinical competence as a theoretical concept, or construct, involves performing several tasks at once–with each informing the others–under substantial time pressure. The standard 15-minute clinical encounter involves the simultaneous use of overlapping skills. Medical students must elicit, in a culturally sensitive manner, key information about the patient, including details about potentially uncomfortable topics like sexuality, mental state, life stressors, addictions, etc. They must conduct a focused physical exam, paying close attention to patient comfort and develop a plan that accounts for the patient's own perspective on his or her own illness. Medical students should educate and counsel when appropriate about behavior changes to improve health outcomes and respond to patient questions in a respectful way all while showing empathy, concern, and respect. All this must be accomplished in a short span of time, often as little as 15 minutes, before meeting with the next patient.

## 1.5 The Objective Structured Clinical Exam (OSCE)

When performance based assessments of clinical skills were introduced, medical educators voiced concern over their objectivity and consistency (Barrows, Williams, & Moy, 1987; Harden, Stevenson, Downie, & Wilson, 1975). For instance, in some assessments, faculty evaluated medical student performance based on observed interactions with a few non-randomly selected patients. Measurement tools designed to capture a holistic impression of medical student performance in these encounters often lacked focus and refinement. Faculty evaluation seemed unstandardized. As a result, measurement of medical student clinical skills seemed either a product of the patient panel or the whim of the faculty member.

These issues directly fueled the development and promotion of the Objective Structured Clinical Examination (OSCE), a performance-based assessment tool designed to promote rigorous, criteria-based, and standardized measures of clinical performance.

In its nascent stages, the OSCE involved simply a series of small, focused clinical cases, or stations, around which medical students rotated (Harden et al., 1975). The student was asked to perform some sort of procedure, like interview a patient, portrayed by an SP, complaining of difficulty breathing, followed by a station in which the student answered questions about his or her findings. In each station, a faculty member completed a score sheet, or checklist, designed to capture medical student completion of specific, discrete, select groupings of task items, as agreed upon earlier by faculty evaluators. These methods allowed medical educators to evaluate medical students on the same predefined, select group of desired clinical behaviors in the same clinical context, all portrayed in a standardized fashion.

Early OSCEs often relied on an expert rater to evaluate medical students in a number of different stations, each involving interaction with an SP. For a number of reasons, including the cost and the limited availability of clinical faculty, expert-rated high-stakes examinations were judged as not feasible. In addition, the push within health care for more patient-centered care, which accounts for and incorporates the patient's perspective, expectations and feelings (Levenstein, McCracken, McWhinney, Stewart, & Brown, 1986), has increased the need for methods of assessing medical student communication and interpersonal skills, rapport building, and professionalism from the patient's vantage point. SPs, rather than expert faculty observers, can oftentimes best evaluate medical student ability to deliver care in a patient-centered manner. This paved the way for the development of methods of scoring

medical student performance using SP ratings. With the USMLE now using medical student patient notes in the construction of scores of student competence, there is new interest in the role of medical student self-report in scoring performance.

## 1.6 Background of this study

Most medical schools in the United States have incorporated OSCEs using SPs into their training programs, fueled in part by the national licensing examination, which includes a component designed to measure medical student clinical skills. Developed by an 8-school consortium of institutions dedicated to the assessment of clinical competence, the multi-station Clinical Performance Examination (CPX) is a clinical assessment tool administered to all medical students across the Consortium in their final, fourth year of training. All medical schools in the consortium share the same patient stations and SP checklists, and attempts are made to standardize SP training across institutions (May, 2008). All medical students are required to take the CPX, rotating through eight 15-minute stations in a half-day, all designed to simulate a routine patient encounter in the clinic. The Consortium designed the examination to both measure medical student clinical skills and prepare students for the USMLE Step 2 CS, a national multi-station standardized patient examination required of all medical students seeking licensure in the United States and completed generally during the fourth year of medical school prior to residency.

In 2012, the Federation of State Medical Boards (FSMB) and the National Board of Medical Examiners (NBME) altered the Step 2 CS in a fundamental way, electing to use data collected from medical student documentation of each patient encounter to partially

construct scores of clinical competence. Whereas scores of student history taking and physical examination had previously been constructed based on SP ratings of student behaviors, scores will now be based partially on evaluations by trained physician raters of the medical student's note. Though likely intended to address concerns associated with the reliability and validity of SP ratings, this new scoring strategy also introduces a new source of potential inaccuracy, that of the medical student and penalizes medical students who correctly perform but incorrectly report on the patient encounter.

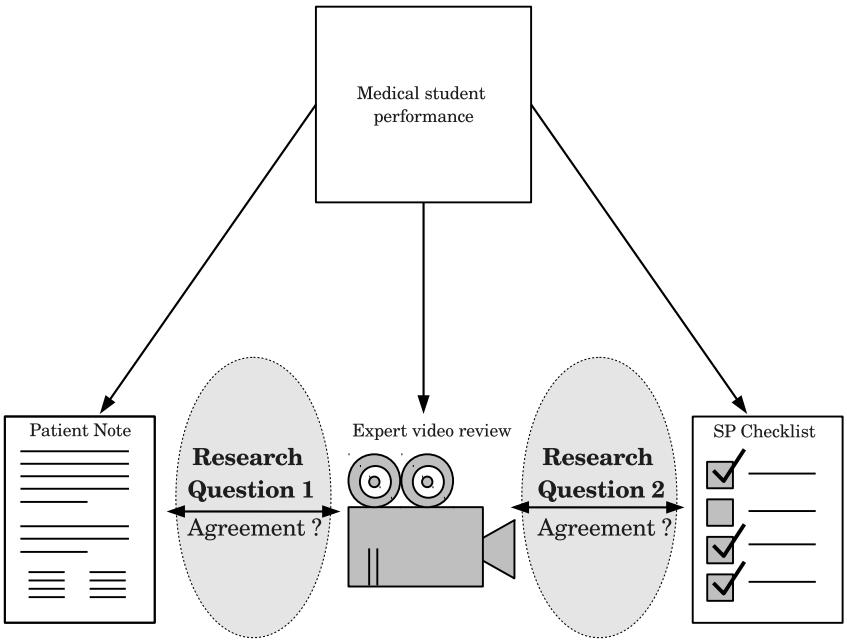
## 1.7 Research questions

This study investigates medical student ability to correctly report performance in a clinical encounter based on documentation in a self-report patient note, while taking into consideration different contextual factors that may impact that ability. Specifically, this study is guided by two main questions, illustrated in Figure 1.1:

1. What is the level of agreement between the medical student self-report and the expert rater documentation of the clinical encounter? That is, how does student self-report compare to the expert rater? To what extent does self-report agreement depend on content and context, features of the examination, or characteristics of the medical student?

2. What is the level of agreement between the SP and the expert rater documentation of medical student performance? Does agreement depend on content and context, features of the examination, or characteristics of the medical student?

Figure 1.1: Study research questions




## 1.8 Implications of present study to the teaching profession

The use of self-report–which includes teacher logs, instructional logs, and time diaries as well as teacher questionnaires and surveys–is common practice in education research and teacher evaluation, including in the making of high-stakes decisions such as promotion and tenure and the allocation of funding. Though there is interest in the use of professional self-report in education, studies have noted the potential for error in the self-report (Ball et al., 1999; Hardebeck et al., 1974; Koziol & Burns, 1986; Newfield, 1980). With little explanation for why professionals have difficulty documenting their behaviors correctly, the

continued use of self-report is in question. This study investigates important methodological concerns related to self-report: 1) professional ability to correctly report performance; and 2) the contextual factors that matter most to the veracity of self-report, specifically: a) the content and context of the information recalled, b) the features of the examination situation recalled, and c) characteristics of the professional doing the recall.

Study findings will guide future research and development of teacher self-report, identifying potential threats to professional ability to report performance (i.e., behaviors/activities, context, or professional characteristics) and enabling the creation of strategies to combat those threats. By better understanding the strengths and limitations of self-report, this study aims to guide future research in education and improve the use of teacher self-report as a tool in teacher and faculty assessment and the ongoing improvement of instructional practices. The implications of potential findings will be discussed in detail in Chapter 6.

## 1.9 Chapter summary

Self-report is one of many tools used in the assessment of professionals. Despite its many advantages, its usefulness is often debated. The medical profession provides a unique opportunity to study self-report, as the practice is well established in the medical profession. This study examined the veracity of medical student self-report by comparing medical student report of performance in a self-report to that of an expert. To ground this comparison, this study also established the level of correct report by SPs, the more established raters of student clinical competence. This study then explored the role of potential contextual factors like content and context, examination features, and characteristics of the professional in

student ability to correctly report performance. Study findings have the potential to guide other professions in the development and use of self-report within their own professional assessment programs.

# Chapter 2

# Literature review

Accuracy of OSCE measurement relies heavily on the ability of raters–whether SPs, medical students, or other observers–to recall specific details of a complex encounter under time pressure. Measurement accuracy, therefore, represents the intersection of memory and recall, human ability and error, and contextual factors of the examination. This chapter begins with a general discussion of how recall error can occur, thereby framing medical student self-report inaccuracy not necessarily as a product of cheating or falsification but as a natural occurrence given the limits of human memory and recall. Then, the chapter reviews accuracy of measurement in OSCEs, highlighting specific studies and identifying gaps in the literature.

## 2.1 Rater memory and recall and analysis of information

Recall accuracy in testing situations, and the occurrence of incorrect report, can be framed by a larger discussion of human recall and the creation of illusory memories, or false memories. In the case of an OSCE, medical educators rely on the ability of raters to correctly recall behaviors immediately following a clinical encounter. Based on studies of human cognition, human recall is imperfect. Even when raters desire to truthfully represent the encounter in the information they provide, the very environment created by the administration of an OSCE can promote poor recall.

Research indicates that recall of conversation is poor (Ross & Sicoly, 1979; Stafford & Daly, 1984). With "egocentric bias" (Ross & Sicoly, 1979) shown in studies of memory of conversation and conversational exchange, subjects more readily recall material they produce, or generate, rather than material they hear. In one study, J. B. Miller, deWinstanley, and Carey (1996) found that conversation partners had better recall of their own contribution to the conversation than that of their partner. J. B. Miller *et al.*. investigated recall error, including recall of ideas that did not appear in the conversation or recall of ideas in the wrong context (e.g., a partner incorrectly attributing an idea to him or herself as opposed to his or her partner), finding that the number of errors was negatively correlated with social skill and positively correlated with social anxiety. In other words, the more socially competent and the less socially anxious an individual, the more correct that individual's recall of both his/her own ideas and that of his/her partner's ideas.

Studies have also shown that an individual's understanding of the world, or schema,

can inhibit story comprehension, thereby producing errors in recall. The effectiveness of a schema is determined by its accessibility, or how easily it comes to mind, and is based on an individual's experience. Kintsch and Greene (1978) found it was hard for individuals to summarize a story that did not fit a familiar story schema. In their study, students were asked to summarize four stories, two Alaskan Indian myths and two stories from the Decameron. Whereas the Decameron follows a European story schema (exposition, complication, and resolution), familiar to most American college students, the Alaskan Indian myths adhered to story conventions culturally specific to those peoples, such as sudden changes in the story's hero and non-chronological and non-causal links between the different parts of the story. Though students did produce summaries that contained correct information about the Alaskan Indian myths they were asked to summarize, their summaries were judged to be less clear about the main events of the myth than their summaries of Decameron stories.

These findings have important implications for raters evaluating performance in an OSCE, be they SPs or medical students. Firstly, research on human cognition indicates that the creation of illusory memories is commonplace in the recall of conversation and that certain factors can contribute to recall ability, such as an egocentric bias and social competence. For both the SP and the medical student, whether or not they themselves generated an idea during the course of the interaction may impact their documentation of medical student behavior. For instance, students may have less difficulty recalling the questions they asked about the patient's symptoms than patient response to physical examination maneuvers. Likewise, SPs may report what they told the student about their illness but have difficulty correctly reporting information shared with them about their diagnosis by the student. Additionally, a medical student's social competence may impact documenta-

tion. In other words, medical student ability to correctly document a clinical encounter may be affected by human ability to participate in a conversation.

Secondly, summarization of a story, parallel to summarization of a clinical encounter can be impacted by how well elements of a story, or features of a clinical presentation, align with a story schema, or a medical student's general schema of a particular illness, also known as an "illness script" (van der Vleuten & Newble, 1995). The experienced clinician builds an illness script around extensive patient experience and medical knowledge (Charlin, Tardif, & Boshuizen, 2000), allowing for efficient processing of clinical information through, essentially, pattern recognition as opposed to the more formal hypothetico-deductive reasoning process taught to medical students. For medical students, who lack experience and (sometimes) knowledge, use of incomplete illness scripts can lead to premature closure, or the tendency to not consider alternate diagnoses after reaching a plausible diagnosis. For instance, medical students may have learned that the classic symptoms of Type I diabetes include frequent urination, extreme thirst and hunger, sudden weight loss, and fatigue. Students may also have seen a couple of patients in the clinic with diabetes who experience some or all of these symptoms. If a medical student learns early in an SP encounter that the patient has some of these symptoms, he or she may be more prone to assume diabetes and to recall only information that supports that diagnosis, for instance recording information about urination, weight loss, and fatigue, and omit other pertinent details from the patient's history that do not fit the student's illness script of diabetes. Likewise, students may even report information they did not actually learn in the encounter, because he or she has created an illusory memory of obtaining certain details that fit the diabetes illness script. In essence, correct report of the clinical encounter by the medical student may be limited by the conventions of human

cognition, memory and recall.

## 2.2 Reliability and validity of OSCE scores

Studies investigating potential threats to the reliability and validity of OSCEs underscore the importance of the control and standardization of the testing environment (e.g., behaviorally-grounded checklist items and SP training) to mitigate measurement error stemming from human subjectivity. Research does in general indicate that SP-rated performance evaluations, when carefully constructed and implemented, are reliable and valid measures of medical student clinical performance. Additional studies have also shown SPs to be highly consistent (Vu & Barrows, 1994) and capable, perhaps even more so than clinical experts (Han, Kreiter, Park, & Ferguson, 2006), of evaluating medical students consistently and accurately on behaviors based on predefined standards.

It is important to note, however, that accepted standards of OSCE score reliability are oftentimes lower in comparison to more traditional knowledge-based examination scores, due to the complex nature of performance based assessment (Newble & Swanson, 1988; Brannick, Erol-Korkmaz, & Prewett, 2011). One major source of concern for OSCE developers is case specificity, or the variation in medical student performance from station to station (van der Vleuten & Swanson, 1990). In fact, in a study of CPX performance scores from 2008 that compared two methods of standard-setting, the authors found that the generalizability of scores (based on SP report) for an 8-station examination was only moderate, $\phi = .48$, for one method of standard-setting, and small, $\phi = .25$, for another (Richter Lagha, Boscardin, May, & Fung, 2012). In both cases, case specificity likely contributed substantially to measurement

error.

## 2.3 SP accuracy in OSCEs

Because OSCEs commonly rely on SPs as evaluators of medical student performance, much of the research in rater accuracy focuses on the SPs ability to correctly recall medical student behaviors following an encounter. Studies have repeatedly shown that trained SPs are capable of evaluating medical students using a checklist with the same accuracy as that of teaching faculty and that there are high levels of agreement between SPs and faculty (Beaulieu et al., 2003; Boulet, McKinley, Norcini, & Whelan, 2002; Luck & Peabody, 2002).

With the advent of the SP as rater, some experts expressed concern over a layperson's ability to accurately rate medical students. Several early studies examined SP ability to report student performance, focusing particularly on the agreement in ratings of student behaviors between the SP and a trained observer, whether a physician or layperson, who used audio or video recordings of encounters to rate medical students (Kopp & Johnson, 1995; Tamblyn, Klass, Schnabl, & Kopelow, 1991; Vu et al., 1992). In most cases SPs correctly rated medical student performance. Similarly, de Champlain et al. (1997) found high rates of agreement between SPs, SP observers, and experienced SP trainers based on analysis of five stations from a larger 12-station SP examination. The authors did identify some select, few items (8.6% of items) with poor agreement, most of which involved detailed physical examination items, which may have been difficult for SP observers and trainers to observe via videotape.

Research has also focused on the relationship between SP documentation and the mea-

surement tools used to collect data. Like de Champlain *et al.*, Vu et al. (1992) did find that SP documentation of medical student performance was good; however, SP ability to correctly report performance was affected significantly by the length of the overall checklist. The authors also noted a significant difference in mean accuracy of report between physical examination items and history items and patient education items. In a subsequent analysis, the authors found that items with low report accuracy shared poor item clarity (e.g., including several behaviors in one item, not explicitly describing an examination maneuver, etc.). Vu *et al.*. also examined the relationship between correct report by SPs and the time of the examination in order to investigate potential SP "fatigue;" they found that SPs correctly reported student performance over the length of the 15-day examination. In one small study of physicians, Luck and Peabody (2002) found high rates of agreement between ratings of physicians made by unannounced SPs and an independent reviewer, with no significant differences in SP ability to correctly report physician performance by the different conditions portrayed by the SPs (i.e., chronic obstructive pulmonary disease, depression, diabetes, or vascular disease) nor by the kind of item, be it history taking, diagnosis, or treatment and management.

One study by de Champlain, MacMillan, Margolis, King, and Klass (1998) investigated the differences in SP ratings of medical student behaviors to explore their potential impact on decisions made regarding student mastery of clinical skills. Though the authors did stress that accuracy of SP scores was not the focus of their study, the study findings do have some interesting implications for the use of SP ratings. The investigators assigned a "pass" to medical students in each of the six cases based on performance on predetermined criteria. In the study, both SPs and SP observers (SPs trained in the stations who were observing

via a monitor in another room), rated medical student performance. Across the different cases, the proportion of medical students classified identically remained quite high, ranging from .86 to .92. Though the authors found little variation in SP-SP observer scores, they did note considerable differences in station difficulty and a high level of case-specificity, meaning that medical student performance varied depending on the station. Like Luck and Peabody (2002), regardless of the station, SPs and SP observers assigned passing scores identically, indicating a high level of accuracy of SP scores in recording medical student behavior.

The overwhelming conclusion from this research is that SPs are capable of correctly rating medical student behaviors. Although some researchers have noted that SP error tends to favor the medical student (de Champlain et al., 1998; Vu et al., 1992), which could lead to false positives, or determining a medical student clinically competent when he or she is not, in fact, competent, most note that given the rigorous training of SPs and the use of well designed data collection tools, this is likely uncommon.

## 2.4 Medical student self-report accuracy in OSCEs

In contrast to investigation of SP ability to report student performance, very little research has examined the ability of medical students to correctly document their performance in the clinical encounter. Recently, Szauter, Ainsworth, Holden, and Mercado (2006) used video review of medical student-SP encounters in three stations to determine the medical student ability to record elements of the physical examination in post-encounter notes. The results were alarming. Szauter *et al.*. found that 96% of notes (199 of 207) contained some level of inaccuracy, or "mismatch", which the investigators classified into different categories,

the most relevant to the present study being under-documentation and over-documentation.

Under-documentation, referred to as "under-report" in the present study, occurs when the medical student neglects to document information from the encounter like, for instance, omitting a physical examination maneuver. Szauter *et al.*. found evidence of under-documentation in 43 % (89 of 207) of the medical student notes. In a separate study that compared information recorded in medical student notes to that found in the SP checklists from a 6-station OSCE, MacMillan, Fletcher, de Champlain, and Klass (2000) observed that medical students under-reported findings from the medical interview, noting an average proportion of "discordance" between the medical student note and the SP checklist of as high as .33 (nearly 3 of 8) for the physical examination items in one station and as high as .31 (roughly 2 of 8) for the history items in another station. Under-report is also observed among practicing physicians, who, for instance, neglect to note in the medical record discussion of preventative care that occurred with the patient (Dresselhaus, Luck, & Peabody, 2002). This has lead researchers to conclude that audit of the medical record may not truthfully represent physician behaviors (Cohen, Ek, & Pan, 2002; Ellis, Blackshaw, Purdie, & Mellsop, 1991).

Hypothesized reasons for omission of important information from the encounter, or "under-report," by medical students include: a) students forgot an action was performed; b) students forgot to record the action; or c) students determined an action was not relevant and/or necessary to document (Szauter et al., 2006). These explanations do not exclude other root causes of error in medical student report, like those outlined above such as egocentric bias, social ability, and story schema.

Over-documentation, referred to as "over-report" in the present study, signifies medical

student report of information in the patient note that was not obtained during the encounter with the SP. Both Szauter et al. (2006) and MacMillan et al. (2000) noted the presence of over-report in medical student notes, though it was observed much less frequently. Both studies mentioned "fabrication" as an obvious cause, though MacMillan *et al.* suggested that a problem with the scoring key used in the study was more likely to blame. Szauter *et al.* went so far as to suggest fraud, a serious crime within health care. In one small study involving 20 practicing physicians (Dresselhaus et al., 2002), the researchers identified over-report rates ranging from 0.098 to 0.397 among individual physicians. The researchers again warned against possible "intentional falsification" by some physicians in the medical record.

Although elaboration, fabrication, and falsification seem easy conclusions to make when encountering over-report in medical student notes, cognitive theory and research, recognizes and affirms the prevalence of illusory memories in human recall. For instance, medical students convinced that a patient is suffering from a particular disease may readily recall supporting information not obtained during the encounter (and likewise omit information that does not support the diagnosis), not to falsify their record or achieve a better grade but because they have formed an inaccurate or incomplete memory.

Based on the limited number of studies investigating medical student self-report, it does appear that medical students do misrepresent the clinical encounter. This may not be so surprising considering the presence of documentation errors in the patient medical record made by practicing physicians (Cohen et al., 2002; Dresselhaus et al., 2002; Ellis et al., 1991; Luck, Peabody, Dresselhaus, Lee, & Glassman, 2000; Peabody, Luck, Glassman, Dresselhaus, & Lee, 2000). Whereas the cause of SP error in report of student performance is often tied

to the item, to the competency domain, or to issues of measurement, like the length of the checklist, the root cause of medical student inaccuracy is not well understood. Though falsification is certainly a concern, as with any examination, there is little evidence to support this explanation. Certainly, there is a need to investigate why medical students are incorrect in their self-reports in order to, if possible, prevent it to ensure proper interpretation of scores.

## 2.5 Relationship between accuracy of the medical student self-report and contextual factors

Many studies investigate potential threats to SP ratings, given that these often form the base of clinical performance evaluation. Attempts to identify potential sources of bias in SP ratings have focused on characteristics like gender and the possible effect of the interaction between SP and student gender on ratings of medical student skills (Chambers, Boulet, & Furman, 2001; Colliver, Vu, Marcy, Travis, & Robbs, 1993; van Zanten et al., 2003). Results from this body of research are mixed. Though some note significant differences in medical student performance scores between male and female SPs, no study has reported a significant interaction between SP and student gender. This implies that although, for example, a female SP may differ in severity from a male SP, she is equally severe or lenient for both male and female medical students.

Few researchers have investigated the relationship between medical student ability to correctly self-report (in the patient note) and the content and context of performance items,

features of examination, and characteristics of the medical student. For instance, to date, no studies have explored how features of the SP-medical student interaction may impact medical student ability to report findings of the encounter. As mentioned previously, parallel research involving practicing physicians suggests that errors are commonplace in the medical record. The majority of research in this area focuses on physician ability to correctly report specific aspects of the patient's medical history, for instance medications or mental health, as compared to self-report by the patient. In essence, though ample evidence demonstrates the inaccuracy of the medical record, few have attempted to relate these errors to the patient-physician encounter, which is a key step in learning how to prevent such errors in the future.

## 2.6 Chapter summary

Studies of human cognition reveal that we are subject to error in the recall of conversation, or the creation of illusory memories. Research has shown that egocentric bias, social anxiety, and story schema can all lead to incorrect recall of details of a conversation or story. Both medical students and SPs are therefore subject to error in recall of the encounter. While much research has indicated that SPs are highly capable of providing accurate ratings of medical student performance, the little research that has investigated the ability of medical students to correctly report information about an encounter in their patient notes has shown medical students oftentimes under- and over-report information. No studies have investigated the role of potential contextual factors in student ability to correctly report performance. It becomes imperative that we better understand medical student self-report as institutions move towards using information furnished by medical students in the scoring of performance

on clinical competence examinations.

# Chapter 3

# Methodology

## 3.1 Overview of study methodology

This study investigated validity of self-report data at the most basic level, examining how truthful is information provided by medical students about their performance in a clinical encounter. Of interest was not simply how well medical students reported behaviors, but also, more importantly, what contextual factors, like, for instance, the gender of the patient, the sequence of the station in the course of the examination, or even features of the encounter itself, affected student ability to correctly report behaviors.

This study considered several important aspects of the medical student self-report. Firstly, this study examined the level of agreement about performance between medical student and expert rater by comparing behaviors reported by medical student to those observed by the expert rater when watching video recordings of those same encounters. Secondly, this study examines how medical student ability to correctly report performance compared to SP ability to correctly report student performance; SPs, after all, are more often used to score

student clinical performance. Finally, this study strove to understand what, if anything, has an impact on medical student ability to correctly report an encounter by considering possible contextual factors. This study is approved by the UCLA Office of Human Research Protection Program.

## 3.2 Settings

Data collected for this study derived from medical student performance on a fourth-year examination at a public medical school in the southwest United States, where medical students are required to successfully pass an 8-station OSCE in order to graduate. Administered at the beginning of the fourth year of medical school, the Clinical Performance Examination (CPX) purports to assess medical student clinical skills.

## 3.3 Participants

Study subjects included fourth year medical students who completed the CPX in 2012. These students had already completed two years of (mostly) non-clinical coursework, such as basic science courses, histopathology, and anatomy, and had just finished a year of clinical rotations through the various sub-specialties like surgery, medicine, pediatrics, psychiatry, etc., as part of the medical team. Of the 183 medical students who participated in the examination, a stratified random sample of 75 students who completed the CPX over the course of three weeks in 2012 was selected based on average student physician-patient interaction (PPI) scores, a measure of social skills (see instruments and scores below for further details),

across the three stations (a subset of the 8 stations found on the examination) that were included in this study.

Documentation of performance included video recordings of student performance, SP documentation, and student self-report of the encounters. Of the 75 students in the study sample, 43% (32 of 75) were female, which was representative of the class. During administration of the examination, 27 (36%) of medical students completed the examination in the first, 34 (45%) in the second, and 14 (19%) in the third partial week. In addition, 31 of the 75 students (41%) completed the examination during a morning session; 44 (59%) completed the examination in the afternoon. On average, student PPI scores (a measure of social skills) were high, averaging 8.6 points out of 10 possible points across the three stations. Video recordings of medical student performance were missing for 3 of 75 students (4%) in station 1 and for 1 of 75 students (1%) in station 2; station 3 had no missing video recordings. Student self-reports were missing for 2 of 75 (3%) students in station 1, for 1 student (1%) in station 2, and for 2 (3%) students in station 3.

## 3.4 Administration of the CPX

The CPX was administered during the first three weeks in June 2012. Medical students were randomly assigned to either morning or afternoon sessions during this period. Medical students arrived at the testing facility to watch a brief instructional video about the CPX. Though students have participated in OSCEs throughout their medical career, the CPX was the longest OSCE they had completed while in medical school, at around five hours. Medical students (and SPs) received a brief break in the middle of each examination session. Over

the course of the examination, the medical students rotated through a series of encounters following a strict time schedule. At the end of the examination, medical students adjourned for a debriefing session with a faculty member.

Prior to the examination, SPs received six hours of training in patient portrayal and in completing the checklist. Two to three different SPs were trained to portray each station. During the examination medical students saw one of these SPs in a given station, meaning that medical students participating in the CPX on a given day likely saw a different panel of SPs than those medical students participating in another session on a different day.

During training, each SP received a detailed station description that provided information about the patient as a person, such as information about his or her professional, social and family life, as well as information about the illness portrayed. This included behaviors such as coughing or pain on movement as well as scripted responses to questions likely to be posed by medical students. One such scripted response was the opening line, as the majority of students ask upon entering the room some variation of, "What brings you into the clinic today?" SPs were trained to provide each student with a standard, scripted response to this question. In station 1 and station 3, these opening lines simply reiterated information posted on the station door outside the encounter ('I'm having some bad back pain' and 'My period won't stop,' respectively). In station 2, the opening line elaborated on the posted information, providing additional details about developing symptoms ("My cough seems to be getting worse. Now I'm experiencing –."). SPs were also instructed to give students the benefit of the doubt when unsure whether or not a student performed a critical action, meaning SP over-reporting was in some respects encouraged.

The CPX consisted of eight 15-minute student-SP encounters, each designed to simulate

a real-life patient encounter in a clinic. Potential diagnoses ranged from muscular strains and fractures to infection to the possibility of a more serious illness like cancer or AIDS. Prior to each encounter, medical students were supplied with basic information about the patient, posted on the door to each encounter. This included the patient's name and reason for visit in the patient's own words (also known as the chief complaint), and some clinical information like the patient's blood pressure and respiratory rate. During each encounter, medical students were instructed to interview the patient, perform an appropriate physical examination, and offer a differential diagnosis and plan, while establishing and maintaining good patient-physician rapport. At the end of each station, the medical student exited the room and the SP completed a checklist using a computer found in each room, indicating "done" or "not done" on a series of performance items as well as rating several items pertaining to student social skills (used to determine student PPI score, described below). In the meantime, the medical students completed a post-encounter activity at small computing workspaces located just outside the rooms. In 3 of the 8 stations, this activity was a student self-report in the form of a patient note, in which the medical student reported the patient's history and physical findings. In addition, the medical student listed for each patient up to three potential diagnoses, in order of likelihood–called the differential diagnosis–and included for each diagnosis justification based on information gathered during the interview. All data, including SP checklists and medical student patient notes, were stored electronically, and all encounters were videotaped using the software program METI Learning Space.

For purposes of examination security, it is not possible to describe patient symptoms in detail, including specific critical action items found on the examination. Table 3.1 details some distinguishing information about the three stations examined in this study.

Table 3.1: Summary of station features

| Station | SP Gender | Age | Severity of portrayed illness | Patient complaint | Supplementary materials provided | Scripted opener | System | Life-threatening illness possible |
|---|---|---|---|---|---|---|---|---|
| 1 | Male | 32 | Acute | Lower back pain | No | Provided no new information from what appeared on the door | Muscular-skeletal | No |
| 2 | Male | 50 | Chronic | Cough | Yes | Elaborated on information provided on door | Cardio-pulmonary | Yes |
| 3 | Female | 44 | Acute | Heavy menses | Yes | Provided no new information from what appeared on the door | Abdominal | Yes |

While all stations were designed to test students on history taking, physical examination, and information sharing skills, there were some important distinctions between the three stations. In station 1, the patient's acute, severe, and sudden onset of lower back pain prompted him to seek medical care. In station 2, the patient's chronic, persistent, lingering cough had worsened, and the student soon learns, thanks to the SP's scripted opening line, exactly how. In station 3, the patient was alarmed by the sudden onset of a heavy menses, prompting her to seek medical care. Interestingly, the patients in stations 1 and 3 do not, in their opening lines, revealed additional information about their symptoms besides what was provided to the students prior to entering the room. In station 2 the SP's scripted opening did reveal additional information about symptom progression critical to proper diagnosis. Though all patients were white adults, they did differ in terms of socio-economic backgrounds, marital and family status, and employment history. There were also important differences among patients with respect to family history (presence of certain diseases), social history (drug and alcohol use), and sexual history.

Another important distinction among stations involved special materials provided to students within the encounter. Students in station 2 were provided with special equipment that augmented elements of the physical examination, producing certain findings when there were, in actuality, none. In addition, in station 3, students who requested specific additional key diagnostic tests were provided with a report of test results by the patient. Correct interpretation of these results in the encounter was key to properly diagnosing the patient and providing appropriate information about next steps. It was also important to note that though patients in all stations were concerned about their symptoms, students were expected in stations 2 and 3, given findings from the history and physical examination, to deliver bad

news, or, in other words, discuss the possibility, following future tests, of a life-threatening illness.

## 3.5 Instruments and Scores

This project involved several data sources, including the behavioral checklist, medical students' self-report (the patient note), the medical student patient-physician interaction (PPI) checklist, and variables related to the encounter.

*The Behavioral Checklist*: The behavioral checklist consisted of a list of behavioral items, predetermined by faculty, that a medical student should perform when confronted with each clinical scenario. Items were organized into three distinct categories: patient history, physical examination, and information sharing, as defined below.

History taking: Information learned during the encounter based on medical student interview of the patient. This included: patient symptoms, progression of illness, past medical history, family medical history, social history (e.g., employment, home situation, drug and alcohol use), sexual history, and other relevant details about the patient and his or her illness. For instance, learning from the patient that "The pain started 9 days ago." Generally, these items were performed first in the clinical encounter.

Physical examination: Physical examination maneuvers performed during the physical examination of the patient. For instance, "Listened to the heart in four places." Although students may have misinterpreted patient response to that maneuver, they still received credit for performing the behavior. Generally, these items were performed following the history taking portion of the encounter but before the information sharing portion of the

encounter.

Information sharing: Information shared with the patient by the medical student about diagnosis and next steps in terms of subsequent steps and possible treatment. For instance, "Recommended a chest x-ray" or "Told the patient he may have shingles." Items pertaining to diagnosis only provided the correct diagnosis in the wording of the item; therefore, students only received credit for sharing information about the diagnosis if that diagnosis was accurate. Generally, these items were performed last in the encounter, just before the student exited the room.

The full behavioral checklist for each station averaged 25 history, physical examination, and information sharing items. This study applied a sub-set of these items, called "critical actions," selected by an expert panel of five faculty members and two facilitators as part of an institutional effort to develop a criterion-referenced standard for the CPX. A two-step process, the "critical actions" approach first asked faculty to identify a key set of checklist items, called critical actions, the performance of which is "critical to ensure an optimal patient outcome and avoid medical error," followed by a second rating process designed to achieve consensus among faculty on the inclusion of these behaviors (Payne et al., 2008). This resulted in an abbreviated checklist of 4-7 items for each station, or so-called "critical actions."

This abbreviated checklist was used in three ways. First, SPs completed the checklist electronically, indicating "Done" or "Not Done" to each item once the medical student had exited the room. Item-level data for each SP in each encounter were entered into the dataset (0="Not Done", 1="Done") using variable names that indicated station number, the domain of the item, the item number, and originator of the data, the particular SP.

Second, the behavioral checklist was applied to the medical student note, recording items reported and not reported in the patient note. Item-level data from the medical student patient note in each encounter were entered into the dataset (0="Not Reported", 1="Reported") using variable names that indicated station number, the domain of the item, the item number, and origin of the data, the medical student.

Third, the expert rater reviewed video recordings of all medical students participating in the CPX, applying the behavioral checklist to medical student behavior in each encounter. Item-level data from the expert rater in each encounter were entered into the dataset (0="Not Observed", 1="Observed") using variable names that indicated station number, the domain of the item, the item number, and origin of the data, the expert rater.

The data schematic found in table 3.2 illustrates the organization of these data by items, domains, stations, and, finally, raters. Though each station contained items unique to that station, all items fell into one of three domains (history, physical examination, and information sharing). Each station included at least one critical action item in each of these three domains. Scores for each item in each domain from all three stations were provided by all three rating sources: the medical student self-report, the SP checklist, and the expert documentation of the video recording of the clinical encounter.

Table 3.2: Organization of data based on application of the behavioral checklist

| | *Rater:* | *Medical student* | | | | | | | | | | | | *SP* | | | | | | | | | | | | *Expert* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Station (s):*[a] | $s_1$ | | | | | $s_2$ | | | | | $s_3$ | | | | $s_1$ | | | | | $s_2$ | | | | | $s_3$ | | | | $s_1$ | | | | | $s_2$ | | | | | $s_3$ | | |
| | *Domain (d):*[b] | $d_1$ | $d_2$ | $d_3$ | | $d_1$ | $d_2$ | $d_3$ | | | $d_1$ | $d_2$ | $d_3$ | | | $d_1$ | $d_2$ | $d_3$ | | $d_1$ | $d_2$ | $d_3$ | | | $d_1$ | $d_2$ | $d_3$ | | | $d_1$ | $d_2$ | $d_3$ | | $d_1$ | $d_2$ | $d_3$ | | | $d_1$ | $d_2$ | $d_3$ | |
| Student | *Item (i):*[c] | $i_1$ | $i_2$ | $i_3$ | $i_4$ | $i_5$ | … | $i_{14}$ | $i_{15}$ | $i_{16}$ | $i_{17}$ | $i_{18}$ | | | $i_1$ | $i_2$ | $i_3$ | $i_4$ | $i_5$ | … | $i_{14}$ | $i_{15}$ | $i_{16}$ | $i_{17}$ | $i_{18}$ | | | $i_1$ | $i_2$ | $i_3$ | $i_4$ | $i_5$ | … | $i_{14}$ | $i_{15}$ | $i_{16}$ | $i_{17}$ | $i_{18}$ | |
| 1 | | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| 2 | | 0 | 1 | … | … | … | … | … | … | … | … | … | … | .. | 0 | | 1 | 0 | … | … | … | … | … | … | … | … | … | … | .. | 0 | | 0 | 0 | … | … | … | … | … | … | … | … | … | .. | 1 |
| . | | 1 | 0 | … | … | … | … | … | … | … | … | … | … | .. | 1 | | 0 | 1 | … | … | … | … | … | … | … | … | … | … | .. | 0 | | 1 | 0 | … | … | … | … | … | … | … | … | … | .. | 1 |
| . | | 0 | 1 | … | … | … | … | … | … | … | … | … | … | … | 1 | | 0 | 1 | … | … | … | … | … | … | … | … | … | … | .. | 1 | | 1 | 0 | … | … | … | … | … | … | … | … | … | .. | 0 |
| . | | 0 | 1 | … | … | … | … | … | … | … | … | … | … | .. | 1 | | 1 | 1 | … | … | … | … | … | … | … | … | … | … | .. | 0 | | 1 | 0 | … | … | … | … | … | … | … | … | … | .. | 1 |
| 75 | | 1 | 0 | … | … | … | … | … | … | … | … | … | … | .. | 1 | | 1 | 1 | … | … | … | … | … | … | … | … | … | … | .. | 0 | | 1 | 0 | … | … | … | … | … | … | … | … | … | .. | 1 |

[a] Refers to station 1 ($s_1$), station 2 ($s_2$), and station 3 ($s_3$)

[b] Refers to history taking ($d_1$), physical examination ($d_2$), and information sharing ($d_3$) item content domains

[c] Refers to items 1 through 18, across the three stations

*Medical Student Self-Report (Patient Note)*: The medical student patient note required medical students to describe the history obtained from the patient, detail the physical examination, and provide a differential diagnosis, or list of potential diagnoses, in order of likelihood, as well as a list of diagnostic studies (e.g., chest x-ray, stress test, etc.) to perform. The expert rater applied the behavioral checklist, as described above, to the medical students' self-reports to determine what critical actions the medical student reported they did perform.

*Medical student Patient-Physician Communication (PPI) Checklist*: In each encounter, SPs evaluated medical student performance on a series of items (0="Not Observed", 1="Observed") related to social skills, specifically communication and the medical student's patient-centeredness during the encounter. SP ratings on these items were entered into the database. Medical students did not self-evaluate their social skills in the encounter nor was it possible to identify these individual elements within the student self-reports (the patient notes), meaning the PPI checklist was not applied to the patient note or the video recordings by the expert rater. Because the PPI score is based on the lived experience of the SP within the encounter, and several of the items do not necessarily involve easily observable behaviors, the PPI score used in this study was determined solely based on SP ratings.

*Contextual Factors: Features/Variables Related to the Encounter*: Contextual factors of interest included: context and content of information, features of the examination, and features of the professional. In addition to those variables defined above, the following variables were also entered into the dataset: the SP gender (0=male, 1=female), the medical student gender (0=male, 1=female), date of the examination (06/04/2012 through 06/20/2012), the session time (morning or afternoon), and the sequence of the encounter in the course of the

examination (first, second, third).

## 3.6 Reliability of scores

Reliability of CPX scores using SP ratings have been shown previously to be low to moderate, depending on the standard used (Richter Lagha et al., 2012), and the accuracy of SP patient portrayal is monitored by SP trainers during administration of the examination (May, 2008). Results of a generalizability study[1] revealed that scores generated by medical students, SPs, and the expert rater report of critical action items exhibited low reliability, as displayed in Table 3.3, likely due to case specificity (van der Vleuten & Swanson, 1990), or variation in performance by students from station to station. Essentially, medical students performed similarly on average, but different medical students performed poorly in different stations.

Scores from all three raters demonstrated low person variance for an examination 8 stations long (the full length of the CPX), particularly for the expert rater, $\sigma_p^2 = .00028$ (8.28% of total variance), and high person-by-station variance, again particularly for the expert rater, $\sigma_{ps,e}^2 = .00208$ (64.54% of total variance). As a result, expert rater scores exhibited rather low reliability, $\phi = .09$, meaning if a different sample of stations were administered to students, very likely different students would be identified as passing and failing the examination. In comparison to the expert rater, scores based on the medical student self-report and SP report of the clinical encounter had relatively higher levels of reliability, $\phi = .33$ and $\phi = .22$, though these values still indicate that if students completed a

[1]For an overview of generalizability theory, please see Brennan, R.L. (2001). *Generalizability Theory.* New York, NY: Springer.

Table 3.3: Reliability of CPX scores

| Source of variance | df | Variance component | 1 station[a] | 3 stations[a] | 8 stations[a] |
|---|---|---|---|---|---|
| | | Medical student | | | |
| Person ($p$) | 71 | $\sigma^2_p$ | 0.00237 (5.85) | 0.00237 (15.72) | 0.00237 (33.14) |
| Station ($s$) | 2 | $\sigma^2_s$ | 0.01583 (37.99) | 0.00528 (35.01) | 0.00198 (27.73) |
| Person-by-station ($ps,e$) | 142 | $\sigma^2_{ps,e}$ | 0.02228 (55.04) | 0.00743 (49.27) | 0.00279 (39.08) |
| Reliability[c] | | | .06 | .16 | .33 |
| | | Standardized patient | | | |
| Person ($p$) | 74 | $\sigma^2_p$ | 0.00066 (3.33) | 0.00066 (9.36) | 0.00066 (21.57) |
| Station ($s$) | 2 | $\sigma^2_s$ | 0.00865 (43.62) | 0.00288 (40.85) | 0.00108 (35.29) |
| Person-by-station ($ps,e$) | 148 | $\sigma^2_{ps,e}$ | 0.01052 (53.05) | 0.00351 (49.79) | 0.00132 (43.14) |
| Reliability[c] | | | .03 | .09 | .22 |
| | | Expert rater | | | |
| Person ($p$) | 70 | $\sigma^2_p$ | 0.00028 (1.12) | 0.00028 (3.27) | 0.00028 (8.28) |
| Station ($s$) | 2 | $\sigma^2_s$ | 0.00818 (32.58) | 0.00273 (31.89) | 0.00102 (30.18) |
| Person-by-station ($ps,e$) | 140 | $\sigma^2_{ps,e}$ | 0.01665 (66.31) | 0.00555 (64.84) | 0.00208 (61.54) |
| Reliability[b] | | | .01 | .03 | .09 |

df = Degrees of freedom.

[a] Variance component (% of total variance)

[b] Reliability calculated using generalizability theory to determine the dependability coefficient ($\phi$), a measure of score generalizability for absolute decisions (like pass/fail decisions) that takes into consideration all sources of measurement error, including station difficulty ($\sigma^2_s$)

different sample of stations, different students would have passed and failed the examination.

## 3.7 Expert qualifications

With 8 years of experience in clinical performance assessment, as the expert rater I am uniquely qualified to record student behavior based on review of video recordings and score medical student patient notes. Firstly, I have been involved in the past in the development of clinical performance examination stations and behavioral checklists and have assisted in the training of SPs. Secondly, the behavioral checklist is intended for use by a non-healthcare provider such as an SP and myself, meaning items are worded for the patient and not a faculty member. Thirdly, I have coded medical student performance based on the behavioral checklist for past projects, including studies of the CPX. A comparison of my scores of medical student patient notes to those of the faculty member revealed similar levels of score reliability, Cronbach's alpha = .21 and .24, respectively. The Pearson correlation coefficient between faculty and expert evaluation of medical student performance of critical action items across all stations was positive, large, and significant, $r(71) = .96$, $p < .001$. There was also no significant difference[2] in the number of students who passed and failed the examination based on my review of the medical student patient notes and faculty review of the patient notes, $\chi^2(1, N = 72) = 0.00$, $p = 1.00$. In essence, there was strong agreement between myself, as the expert, and the faculty member assigned to score patient notes as part of the examination.

[2]McNemar's test statistic calculated using the continuity correction to determine if data exhibit marginal homogeneity. The test statistic was compared to the $\chi^2$ distribution with 1 degree of freedom.

## 3.8 Data analysis

### 3.8.1 Research Question 1

*What is the level of agreement between the medical student self-report and the expert rater documentation of the clinical encounter? That is, how does student self-report compare to the expert rater? To what extent does self-report agreement depend on the content and context of information, features of the examination, or characteristics of the professional?*

To analyze level of agreement between the medical student and the expert rater of the clinical encounter, I focused on the match between the medical student's report and the expert rater's report. The primary variables created included the proportion of checklist items done/not done that matched, or were "in agreement," between the medical student and the expert rater, the proportion of items that were under-reported (behavior was observed by the expert rater but not reported by the the medical student), and the proportion of items that were over-reported (behavior was not observed by the expert rater but reported by the medical student). These proportions were calculated for all items, as well as for each of the three stations and the three domains.

Quantitative analyses, including nonparametric tests for comparing dependent samples, analysis of variance (ANOVA), correlational analysis, and $t$-tests were performed to determine whether level of agreement differed by content and context of information (i.e., station, domain), features of the encounter (i.e., examination date and time, station order), and characteristics of the professional (i.e., medical student gender, SP gender, medical student social skills (PPI) score).

### 3.8.2 Research Question 2

*What is the level of agreement between the SP and the expert rater documentation of medical student performance? Does agreement depend on the content and context of information, features of the examination, or characteristics of the professional?*

To analyze the agreement between the SP and expert rater, I focused on the match between the SP report and the expert rater's report. I created analogous variables to those described above in terms of SP under-reporting and over-reporting (compared to the expert rater), and conducted quantitative analyses, including nonparametric tests for comparing dependent samples, analysis of variance (ANOVA), correlational analysis, and $t$-tests to determine whether the correct report of medical student performance by SPs depends on content and context of information (i.e., station, domain), features of the encounter (i.e., examination date and time, station order), and characteristics of the professional (i.e., medical student gender, SP gender, medical student social skills (PPI) score).

## 3.9 Potential significance of findings

This study contributes to our understanding of the accuracy of scores measuring professional competence using self-reports, which is significant to training, evaluation, and promotion in any profession. Educators are exploring the use of student self-report of the encounter (captured in the patient note) as an indicator of medical student ability to effectively gather information from the patient during the medical interview. Scoring of medical student performance may increasingly depend on the student self-report of the medical encounter rather than SP report of the encounter. This strategy, however, poses a potential threat to the va-

lidity of scores and, therefore, the interpretation of medical student clinical competence. If so, such results would call into question the ability to make meaningful decisions about medical student clinical competence from the student's documentation of the encounter in the patient note. This study aims to better understand the strengths and limitations of medical student self-report as a source of information about the encounter, and, consequently, the validity of results for assessment of medical student's clinical competence.

This study also speaks to the accuracy of information provided by self-reports within other professions (e.g. by teachers) in two important ways. Firstly, it contributes to our understanding of the quality of information provided by professionals about their practices (in this study, referred to as level of agreement), and secondly, to our understanding of how this matters to determining competence. Results from my study can be applied to a more general education setting where a practitioner within a field (the teacher versus the medical student/physician) must appropriately recall and report information from a professional context (the classroom versus the examination room) for the purpose of competence assessment (teacher effectiveness versus clinical competence).

This study investigated potential contributors to medical student misrepresentation of behaviors when reporting information about an encounter in the patient note, or the professional self-report. Are low levels of agreement between medical student self-reported behaviors in the patient note and expert ratings of observed behaviors due to station content? Or to the unique combination of medical student and SP? Are female medical students better at reporting correctly information when dealing with a female SP? Can faulty information provided by the medical student in the self-report be explained in part by the quality of interaction between SP and medical student? Investigation of self-report in other professions,

such as among teachers, must consider similar issues. What accounts for discrepancies between teacher self-reported behaviors and observation by students, peers, supervisors, even independent observers? Are these discrepancies a product of the instrument used to collect self-reported practices? Are they related to contextual features of the classroom? Can incorrect self-report of information by teachers be connected to the quality of interaction between teachers and students? Do different approaches to teaching, or different teaching styles, relate to increased teacher ability to accurately document practices. For instance, does adopting the role of facilitator, where student participation and collaboration in the learning process is encouraged, as opposed to one of formal authority, or classroom lecturing where students actively participate less, promote better quality information in a teacher self-report?

Research on the correspondence between self-reports and observation by students, peers, supervisors, and independent observers of behaviors suggests a moderate to low correlation. Professional self-report does have some advantages, namely it is oftentimes less resource-intensive than observation, and it can better capture the professional's own reasoning. In order to improve self-report, it is therefore important to understand what factors might contribute to this low correspondence. What situations, if any, promote improved agreement between self-report and observer ratings? This study tackled these very issues in one context, medical assessment; however, the results can inform practice among other professionals and professionals-in-training, like teachers. Results from my study improve understanding of self-report and can inform assessment design so as to allow self-reports to serve as valid indicators of information otherwise obtained by observation. With this knowledge, researchers can design tools, protocols, and methods to ensure those conditions in data collection. This

potentially can greatly assist current research efforts in the educational research setting by establishing best practices when dealing with self-report.

## 3.10 Chapter summary

This study employed a variety of quantitative methods to investigate, broadly, the impact of using professional self-report on the determination of competence. A stratified random sample of 75 medical students who completed the CPX, a routine clinical competence examination administered to all fourth year medical students, was selected from the larger population of 183 students. Performance data included: a) performance of items found on the behavioral checklist based on SP report, medical student self-report, and expert rater observation; and b) various contextual factors regarding the context and content of performance information (e.g., station, item domain), features of examination (e.g., date and time), and characteristics of the professional (e.g., gender, PPI score). Analyses of these data, involved nonparametric tests for dependent samples, ANOVA, correlations, and $t$-tests to determine: 1) whether medical student and expert rater report of student performance differed significantly, and if so, by what contextual factors; 2) whether SP (the more common source of performance scores for medical students) and expert rater report of medical student performance differed significantly, and if so, by what contextual factors. This study capitalized on the fact that the practice of self-report within medicine is a common practice, providing a unique opportunity to investigate professional self-report in one context with implications for other disciplines like education and teaching.

# Chapter 4

# Results

For any high-stakes examination of professionals, score accuracy is of grave concern. Evaluators are always looking for ways to improve upon their ability to truthfully capture and subsequently score performance. In the medical field, performance examinations have become a standard means of assessing professional competence. Over the decades, there has been a slow but steady progression from use of faculty to the use of standardized patients (SPs) to rate student performance. New techniques are now being developed, however, that rely on the medical student self-report of the encounter to generate performance scores by, most notably, the National Board of Medical Examiners (NBME) on their own clinical performance examination, the United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills. Given the widespread influence of the USMLE, it is likely that these changes to how scores are constructed will be mimicked in other medical institutions. However, use of student self-report in making decisions regarding professional competence raises questions as to the ability of medical students to truthfully capture their own performance in a clinical encounter. How well do students report performance of an encounter in the self-report as

compared to an expert observer? How does student level of agreement with the expert compare to the more commonly used rater, the SP's, level of agreement with the expert rater? Are students just as correct in their report of performance as SPs, or are they worse? These are important questions to answer for any institution considering the use of medical student self-report to score clinical competence.

This chapter highlights the difference in reporting performance on the CPX based on the different sources of ratings: the expert rater, the medical student, and the SP. This chapter begins by examining the consequences of scoring performance based on the different rating sources, specifically its impact on determination of medical student competence. Then the chapter will tease apart these differences in scoring by firstly presenting overall performance based on the expert rater's observation of medical student performance, and then, secondly, by comparing both medical student self-report and SP documentation to that of the expert rater. Analyses investigating agreement focused on proportion of critical action items accurately reported, over-reported, and under-reported.

## 4.1 Consequences of scoring performance based on different rating sources

Perhaps the most important issue underlying this study is whether or not it ultimately matters on whose information we rely to score medical student performance. Results suggested that the source of information about performed behaviors does affect overall performance scores on the examination.

In order to pass the examination, students were expected to pass all three stations by performing a minimum number of critical action items in each station. As can be seen in Table 4.1, the pass rate was 18% (13 of 73) of students based on student self-report, 68% (51 of 75) of students based on the SP report, and 42% (30 of 72) of students based on the expert rater's observations.

Table 4.1: Medical students who passed the examination by rating source

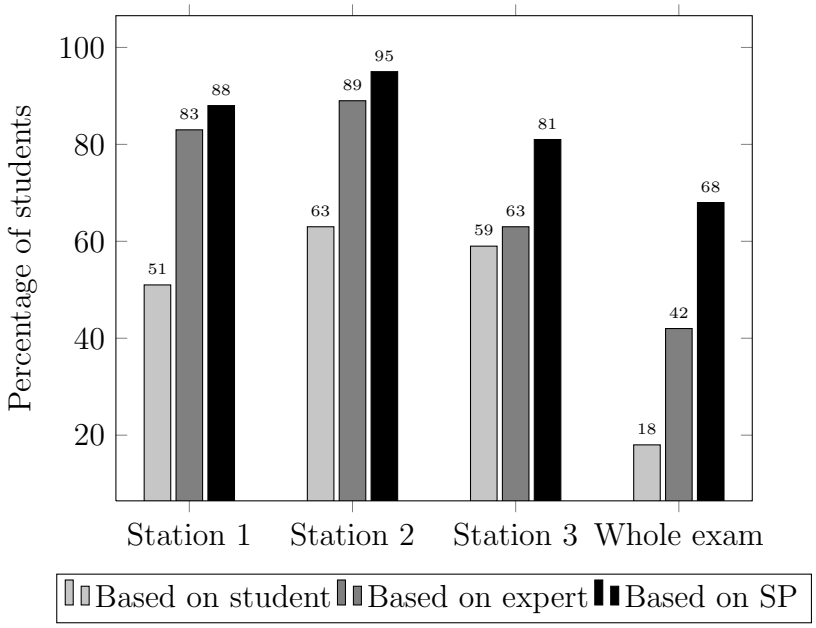
| | | | No. of Students (%) Who Passed | | |
|---|---|---|---|---|---|
| Station | No. of CI | Minimum no. CI to pass | Based on SP (checklist) $N = 75$ | Based on student (self-report) $N = 73$ | Based on expert (observation) $N = 72$ |
| 1 | 7 | 5 | 66 (88) | 37 (51) | 60 (83) |
| 2 | 4 | 4 | 71 (95) | 46 (63) | 64 (89) |
| 3 | 7 | 6 | 61 (81) | 43 (59) | 45 (63) |
| Whole exam | | | 51 (68) | 13 (18) | 30 (42) |

CI = Critical items.
*Note.* To pass the examination, students had to perform the minimum number of critical items in every station.

Figure 4.1 further illustrates the disparity in the proportion of students who passed the examination between the different raters. Student under-report of performance lead to much lower pass rates based on medical student self-report in each of the stations when compared to scores based on expert rater documentation. SP over-report lead to inflated performance scores when compared to the expert. By stipulating that students had to pass all three stations to pass the examination as a whole, incorrect report by both the medical student and the SP grossly exaggerated the proportion of students who passed and failed the examination.

A McNemar's test comparing dependent proportions indicated a significant difference in the proportion of medical students who passed the examination based on the SP perfor-

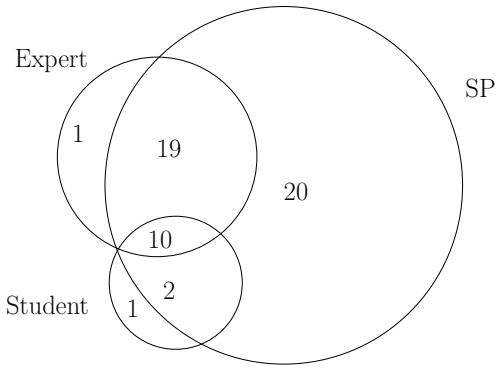
Figure 4.1: Percentage of medical students who passed the examination by rater




mance score as compared to the expert rater, $p = .017$. A McNemar's test also indicated a significant difference in the proportion of medical students who passed the examination based on the medical student self-report as compared to the expert rater, $p = .001$. Whether or not a medical student passed the examination, then, depended on whose documentation of performance was used to generate scores, indicating that real differences in rating sources do exist, that these differences are consequential and can have a serious impact on the determination of student competence, while also underscoring the need to determine potential sources of disagreement to improve the validity of data.

Figure 4.2 displays the overlap in students who passed the examination based on scores by the three different raters: expert, medical student, and SP.

Figure 4.2: Number of students who passed the examination by rater




Of the 75 students included in the study sample, 53 (71%) passed the examination based on performance scores by at least one rater; only 10 students were identified by all three raters as clinically competent. Of the 52 individual medical students identified as having passed the examination by the expert, the SP or both, only 29 students (56%) passed the examination based on both expert and SP scores. Medical student agreement with the expert was even worse, with only 10 of 33 (30%) individual medical students identified as having passed the examination by both the expert and the medical student. So while both the medical students and the SPs were oftentimes in disagreement with the expert rater, medical student disagreement was substantially worse.

## 4.2 Description of medical student performance

Table 4.2 displays medical student performance based on expert observation of the encounters. For easy reference, Table 4.2 also displays performance scores by item from the other two rating sources referenced in this study– the medical student and the SP–allowing for easy comparison between the different rating sources. In general, medical student per-

formance on each of the critical action items was quite high with, on average, students performing 84% of each critical action item. While medical students reported only 75% of each critical action item, SPs reported 88% of each critical action item. These values highlight a general theme, discussed in greater detail below, that medical students tended on average to under-report while SPs over-reported performance.

Table 4.2: Medical students clinical performance by item by rater source

| | | | No. of Students (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Observed by expert | | | Reported by student | | | Reported by SP | | |
| Station | Item | Content domain[a] | *N* | Performed | Not performed | *N* | Reported | Not reported | *N* | Reported | Not reported |
| 1 | 1 | History | 72 | 71 (99) | 1 (1) | 73 | 71 (97) | 2 (3) | 75 | 74 (99) | 1 (1) |
| 1 | 2 | History | 72 | 57 (79) | 15 (21) | 73 | 49 (67) | 24 (33) | 75 | 68 (91) | 7 (9) |
| 1 | 3 | History | 72 | 26 (36) | 46 (64) | 73 | 15 (21) | 58 (80) | 75 | 29 (39) | 46 (61) |
| 1 | 4 | History | 72 | 53 (74) | 19 (26) | 73 | 35 (48) | 38 (52) | 75 | 69 (92) | 6 (8) |
| 1 | 5 | Physical | 72 | 56 (78) | 16 (22) | 73 | 41 (56) | 32 (44) | 75 | 38 (51) | 37 (49) |
| 1 | 6 | Physical | 72 | 64 (89) | 8 (11) | 73 | 45 (62) | 28 (37) | 75 | 67 (89) | 8 (11) |
| 1 | 7 | Information | 72 | 71 (99) | 1 (1) | 73 | 73 (100) | 0 (0) | 75 | 74 (99) | 1 (1) |
| 2 | 1 | History | 74 | 74 (100) | 0 (0) | 74 | 72 (97) | 2 (3) | 75 | 75 (100) | 0 (0) |
| 2 | 2 | History | 74 | 71 (96) | 3 (4) | 74 | 65 (88) | 9 (12) | 75 | 75 (100) | 0 (0) |
| 2 | 3 | Physical | 74 | 74 (100) | 0 (0) | 74 | 54 (73) | 20 (27) | 75 | 75 (100) | 0 (0) |
| 2 | 4 | Information | 74 | 67 (91) | 7 (10) | 74 | 74 (100) | 0 (0) | 75 | 71 (95) | 4 (5) |
| 3 | 1 | History | 75 | 75 (100) | 0 (0) | 73 | 72 (99) | 1 (1) | 75 | 75 (100) | 0 (0) |
| 3 | 2 | History | 75 | 67 (89) | 8 (11) | 73 | 62 (85) | 11 (15) | 75 | 65 (87) | 10 (13) |
| 3 | 3 | History | 75 | 70 (93) | 5 (7) | 73 | 55 (75) | 18 (25) | 75 | 73 (97) | 2 (3) |
| 3 | 4 | Physical | 75 | 72 (96) | 3 (4) | 73 | 65 (89) | 8 (11) | 75 | 72 (96) | 3 (4) |
| 3 | 5 | Physical | 75 | 73 (97) | 2 (3) | 73 | 59 (81) | 14 (19) | 75 | 73 (97) | 2 (3) |
| 3 | 6 | Information | 75 | 38 (51) | 37 (49) | 73 | 44 (60) | 29 (40) | 75 | 65 (87) | 10 (13) |
| 3 | 7 | Information | 75 | 45 (60) | 30 (40) | 73 | 54 (74) | 19 (26) | 75 | 49 (65) | 26 (35) |
| Average total (%) performance[b] | | | 71 | 15.2 (84.3) | 2.8 (15.7) | 73 | 13.6 (75.4) | 4.4 (24.4) | 75 | 15.7 (88.0) | 2.3 (12.0) |

*Note.* Due to rounding, row percentages may not sum to 100.

[a] Refers to history taking, physical examination, and information sharing item content domains as described in Chapter 3

[b] Average number of items out of 18.

Some critical action items were performed by 100% of students; other items were performed by a smaller proportion of students, as low as 36%. Differences appeared across stations (Stations 1, 2, 3) and across item domains (history, physical examination, information sharing). To test differences in performance (again, based on expert observation) across stations, a repeated measures analysis of variance (ANOVA) was conducted. There was a statistically significant effect of station on the proportion of critical action items performed, $F(2, 140) = 39.06$, $p < .01$. Pairwise comparisons revealed a statistically significant difference in performance of critical action items by students between station 1 and station 2, $p < .01$, and also between station 2 and station 3, $p < .01$. Station 2, in which a 50 year old male patient presented with a bad cough, was significantly less difficult for students, whereas station 1, in which a 32 year old male patient complained of lower back pain, proved the most difficult.

Though all stations were designed to simulate an out-patient visit in a clinical setting, there were some important discriminating features between the three stations, summarized again in Table 4.3, that could explain why students found some stations, like station 1, so difficult.

Table 4.3: Summary of distinguishing station features

| Station | SP Gender | Severity of portrayed illness | Supplementary materials provided | Scripted opener | System | Severity of potential diagnoses |
|---|---|---|---|---|---|---|
| 1 | Male | Acute | None | Provided no new information from what appeared on the door: "I'm have bad back pain" | Muscular-skeletal | Likelihood of life-threatening illness low |
| 2 | Male | Chronic | Equipment to enhance physical examination findings provided to all students | Elaborated on information provided on door: "My cough seems to be getting worse. Now I'm experiencing –." | Cardiopulmonary | Diagnosis included potential life-threatening illness |
| 3 | Female | Acute | Physical examination findings provided only if student asked | Provided no new information from what appeared on the door: "My period won't stop." | Abdominal | Diagnosis included potential life-threatening illness |

Firstly, it is possible that these differences in performance indicate a weakness in the medical school curriculum. Perhaps students were simply less familiar with how to approach and evaluate the patient's lower back pain in station 1. Another possibility is that scripted differences between the patients created more or less difficult encounters for the medical students. For instance, in station 2, the patient quickly divulged important information about the progression of his cough to the medical student as part of his scripted opening. In station 1 and 3, the opening lines were more ambiguous, with the patient reiterating only the information that the student read prior to entering the room. This scripted opener may have allowed students to more readily create a productive line of inquiry in station 2. Still a third possibility is that special materials provided to students in station 2 and 3 gave them an advantage in those encounters over station 1. In station 2, special equipment that augmented findings of the physical examination was provided to students whereas in station 3, in which a 44 year old female patient complained of a heavy menses, students obtained the results of additional tests. Use of these additional materials was key to the proper diagnosis of these patients. Station 1, however, had no such materials, perhaps making the station more ambiguous for medical students and therefore more difficult to navigate.

A parallel analysis examined differences in performance (again, based on expert observation) across domains. There was a statistically significant effect of domain on the proportion of critical action items performed, $F(1.59, 110.98) = 23.88$, $p < .01$.[1] Pairwise comparisons revealed a statistically significant difference in proportion of critical action items performed by students between history and physical examination domains, $p = .001$, history and in-

[1]Mauchly's test indicated the assumption of sphericity had been violated, $\chi^2(2) = 23.14$, $p < .01$. Degrees of freedom were therefore corrected using Huynh-Feldt estimates of sphericity ($\epsilon = 0.79$).

formation sharing domains, $p = .003$, and between physical examination and information sharing domains, $p = .003$. Physical examination was the least difficult domain for students, whereas information sharing proved the most difficult.

Training and experience could explain why students had less difficulty performing physical examination items and more difficulty performing information sharing items. One plausible explanation for why students had less difficulty performing physical examination items than history and information sharing items is that students were trained to perform complete physical examinations of each system (e.g., cardiovascular examination, pulmonary examination, abdominal examination, etc.). Therefore, as long as the student initiated an examination of the proper system, it was likely that he or she would perform a majority of individual examination items, including the critical physical examination items. Whether or not the student understood the importance of these items and was able to correctly interpret the results, however, cannot be captured by watching the video recording of the encounter. Obtaining credit for information sharing items, on the other hand, required medical students to process and interpret in the moment the patient's history and physical examination. It is also possible that information sharing is the domain in which most medical students lack real-world experience, as this portion of the patient encounter is likely handled by an attending physician and not the medical student.

## 4.3 Medical student agreement with the expert rater

Table 4.4 displays agreement between what the medical student reported doing and what the expert rater observed the medical student doing in the encounter. On average,

medical students were in agreement with the expert rater about performance on nearly 83% of critical action items. Sources of disagreement stemmed substantially more from medical student under-reporting (12.1% of critical action items) than over-reporting (4.2% of critical action items). That is, in instances of disagreement, medical students more often failed to report behavior thy were observed performing than visa versa.

Table 4.4: Level of agreement between medical student and expert rater of performance by item

| | | | | No. of Students (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | In Agreement | | In Disagreement | | | | |
| Station | Item | Content domain[a] | $N$ | Observed by expert, reported by student | Not observed by expert, not reported by student | Not observed by expert, reported by student (over-report) | Observed by expert not reported by student (under-report) | Total in agreement | Total in disagreement | Test statistic[b] |
| 1 | 1 | History | 71 | 68 (96) | 0 (0) | 1 (1) | 2 (3) | 68 (96) | 3 (4) | 0.00 |
| 1 | 2 | History | 71 | 42 (59) | 9 (13) | 6 (8) | 14 (20) | 51 (72) | 20 (28) | 2.45 |
| 1 | 3 | History | 71 | 11 (15) | 41 (58) | 4 (6) | 15 (21) | 52 (73) | 19 (27) | 5.26 |
| 1 | 4 | History | 71 | 33 (46) | 17 (24) | 1 (1) | 20 (28) | 50 (70) | 21 (29) | 15.43* |
| 1 | 5 | Physical | 71 | 40 (56) | 15 (21) | 1 (1) | 15 (21) | 55 (77) | 16 (22) | 10.56* |
| 1 | 6 | Physical | 71 | 44 (62) | 8 (11) | 0 (0) | 19 (27) | 52 (73) | 19 (27) | 17.05* |
| 1 | 7 | Information | 71 | 70 (99) | 0 (0) | 1 (1) | 0 (0) | 70 (99) | 1 (1) | – |
| 2 | 1 | History | 73 | 71 (97) | 0 (0) | 0 (0) | 2 (3) | 71 (97) | 2 (3) | – |
| 2 | 2 | History | 73 | 63 (86) | 2 (3) | 7 (10) | 1 (1) | 65 (89) | 8 (11) | 3.13 |
| 2 | 3 | Physical | 73 | 54 (74) | 0 (0) | 0 (0) | 19 (26) | 54 (74) | 19 (26) | – |
| 2 | 4 | Information | 73 | 66 (90) | 0 (0) | 7 (10) | 0 (0) | 66 (90) | 7 (10) | – |
| 3 | 1 | History | 73 | 72 (99) | 0 (0) | 0 (0) | 1 (1) | 72 (99) | 1 (1) | – |
| 3 | 2 | History | 73 | 59 (81) | 5 (7) | 3 (4) | 6 (8) | 64 (88) | 9 (12) | 0.44 |
| 3 | 3 | History | 73 | 55 (75) | 5 (7) | 0 (0) | 13 (17) | 60 (82) | 13 (17) | 11.08* |
| 3 | 4 | Physical | 73 | 65 (89) | 3 (4) | 0 (0) | 5 (7) | 68 (93) | 5 (7) | 3.20 |
| 3 | 5 | Physical | 73 | 59 (81) | 2 (3) | 0 (0) | 12 (16) | 61 (84) | 12 (16) | 10.08* |
| 3 | 6 | Information | 73 | 28 (38) | 21 (29) | 16 (22) | 8 (11) | 49 (67) | 24 (33) | 2.04 |
| 3 | 7 | Information | 73 | 38 (52) | 14 (19) | 16 (22) | 5 (7) | 52 (71) | 19 (29) | 4.76 |
| Average total (%) performance[c] | | | | 12.9 (71.5) | 2.0 (11.2) | 0.8 (4.2) | 2.2 (12.1) | 14.9 (82.7) | 2.9 (16.3) | |

*$p < .003$.

*Note.* Due to rounding, row percentages may not sum to 100.

[a] Refers to history taking, physical examination, and information sharing item content domains as described in Chapter 3.

[b] McNemar's test statistic calculated using the continuity correction to determine if data exhibit marginal homogeneity. The test statistic was compared to the $\chi^2$ distribution with 1 degree of freedom.

[c] Average number of items out of 18.

Based on a series of McNemar's tests comparing the dependent proportions of students who documented performing a critical action item to those who were observed performing a critical action item by the expert rater, statistically significant differences at the $p < .003$ level ($p$-value adjusted using Bonferroni correction) existed between the medical student self-report and the expert documentation of the encounter on the performance (or non-performance) of 5 items (see Table 4.4): station 1, item 4; station 1, item 5; station 1, item 6; station 3, item 2; and station 3, item 5. These items, found in stations 1 and 3, spanned the history and physical examination domains, and all displayed the highest levels of under-reporting by medical students. Reasons for disagreement between the medical students and the expert rater were not immediately clear.

### 4.3.1 Agreement between medical student and expert rater by content and context of information

Contextual factors of interest included those related to the content and context of information collected, specifically differences potentially attributable to station (station 1, 2, and 3) and item content domain (history, physical examination, and information sharing). Analyses first addressed differences by station in levels of agreement between medical student and expert using repeated measures ANOVA, then differences in disagreement, specifically under- and over-reporting, using doubly multivariate repeated measures ANOVA.[2] Parallel analyses examine differences in levels of agreement and levels of disagreement by item content domain.

[2]Though measures of agreement between the medical student and the expert rater were not normally distributed, use of non-parametric statistical techniques produced results similar to those reported here.

Table 4.5 displays agreement between the medical student and expert rater by station. On average, medical students were in agreement with the expert rater on 83% of items in a station. Once again, under-reporting was more prevalent than over-reporting on average in stations (12% of items versus 4% of items, respectively).

Table 4.5: Level of agreement between medical student and expert rater of performance by station

| | | | *M* (*SD*) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | In Agreement | | In Disagreement | | | |
| Station | CI | *N* | Observed by expert, reported by student | Not observed by expert, not reported by student | Not observed by expert, reported by student (over-report) | Observed by expert, not reported by student (under-report) | Total in agreement | Total in disagreement |
| | | | No. of items | | | | | |
| 1 | 7 | 71 | 4.3 (1.2) | 1.3 (1.0) | 0.2 (0.4) | 1.2 (0.9) | 5.6 (1.0) | 1.4 (1.0) |
| 2 | 4 | 73 | 3.5 (0.6) | 0.0 (0.2) | 0.1 (0.3) | 0.4 (0.6) | 3.5 (0.6) | 0.5 (0.6) |
| 3 | 7 | 73 | 5.2 (1.2) | 0.7 (0.9 | 0.5 (0.7) | 0.7 (0.9) | 5.8 (1.1) | 1.2 (1.1) |
| Average station (*SD*) performance[a] | | | 4.2 (0.7) | 0.7 (0.5) | 0.3 (0.3) | 0.7 (0.5) | 5.0 (0.6) | 1.0 (0.6) |
| | | | Proportion of items | | | | | |
| 1 | 7 | 71 | 0.62 (0.17) | 0.18 (0.14) | 0.03 (0.06) | 0.17 (0.13) | 0.80 (0.14) | 0.20 (0.14) |
| 2 | 4 | 73 | 0.87 (0.16) | 0.01 (0.04) | 0.03 (0.08) | 0.10 (0.14) | 0.88 (0.16) | 0.12 (0.16) |
| 3 | 7 | 73 | 0.73 (0.17) | 0.10 (0.13) | 0.07 (0.10) | 0.10 (0.13) | 0.83 (0.15) | 0.17 (0.15) |
| Average station (*SD*) performance[b] | | | 0.74 (0.11) | 0.10 (0.07) | 0.04 (0.05) | 0.12 (0.09) | 0.83 (0.10) | 0.16 (0.10) |

CI = Critical items.

*Note.* Due to rounding, row percentages may not sum to 100.

[a] Average number of items by station.

[b] Average proportion of items by station.

Medical students had the highest level of agreement with the expert rater in station 2 (the least difficult station, see Table 4.2). In contrast, in station 1 students under-reported 17% of items. That is, on average they performed but failed to document 17% of critical action items in station 1 (back pain). Station 3 (heavy menses) had the highest incidence of over-reporting (7%). That is, on average, students reported performing 7% of the critical action items that the expert rater did not observe them performing in the encounter.

*Differences in agreement by station.* To test differences in agreement between the medical student and the expert rater among stations, a repeated measures ANOVA was performed. There was a statistically significant effect of station on the proportion of critical action items reported appropriately by the medical student, $F(2, 136) = 4.80$, $p = .010$. Pairwise comparisons revealed a statistically significant difference in the level of agreement between student self-report and expert rater documentation between station 1 and station 2, $p = .007$.

*Differences in over-reporting by station.* To test differences in the nature of disagreement between the medical student and the expert rater among stations, specifically student propensity to over- and under-report critical action items, a doubly multivariate repeated measures ANOVA was employed. The results indicated a statistically significant effect of station on the proportion of critical action items over-reported, $F(1.78, 120.72) = 7.70$, $p = .001$. [3] Pairwise comparisons revealed a significant difference in over-reporting between station 1 and station 3, $p = .004$, and between station 2 and station 3, $p = .014$. In both in-

[3]Mauchly's test indicated that the assumption of sphericity for proportion over-reported had been violated, $\chi^2(2) = 11.16$, $p = .004$. Degrees of freedom were therefore corrected using Huynh-Feldt estimates of sphericity ($\epsilon = 0.89$).

stances, students on average over-reported more in station 3, the station involving a patient experiencing an unusually heavy menses, than in the other two stations.

*Differences in under-reporting by station.* The results also indicated a statistically significant effect of station on the proportion of critical action items under-reported, $F(2, 136) = 8.04$, $p = .001$. Pairwise comparisons revealed a significant difference in under-reporting between station 1 and station 2, $p = .002$, and between station 1 and station 3, $p = .003$. Students on average under-reported more in station 1, the station involving a patient complaining of back pain, than in the other two stations.

Parallel analyses examining the effect of domain on proportion of critical action items correctly reported, over-reported, or under-reported were also conducted. Table 4.6 displays the agreement between medical student and expert rater by domain. On average, students were in agreement with the expert rater on 82% of critical action items in each domain.

Table 4.6: Level of agreement between medical student ($N$=69) and expert rater by domain

| | | $M$ ($SD$) | | | | | |
|---|---|---|---|---|---|---|---|
| | | In Agreement | | In Disagreement | | | |
| Domain[a] | CI | Observed by expert, reported by student | Not observed by expert, not reported by student | Not observed by expert, reported by student (over-report) | Observed by expert, not reported by student (under-report) | Total in agreement | Total in disagreement |
| | | No. of Items | | | | | |
| History | 9 | 6.5 (1.2) | 1.1 (0.9) | 0.2 (0.4) | 1.0 (1.0) | 7.6 (1.3) | 1.3 (1.2) |
| Physical | 5 | 3.6 (1.0) | 0.4 (0.6) | 0.0 (0.1) | 0.9 (0.8) | 4.0 (0.8) | 1.0 (0.8) |
| Information | 4 | 2.7 (0.9) | 0.5 (0.8) | 0.5 (0.7) | 0.2 (0.4) | 3.2 (0.7) | 0.8 (0.7) |
| Average domain ($SD$) performance[b] | | 4.1 (0.6) | 0.8 (0.5) | 0.3 (0.3) | 0.7 (0.5) | 4.9 (0.6) | 1.0 (0.6) |
| | | Proportion of Items | | | | | |
| History | 9 | 0.66 (0.14) | 0.18 (0.14) | 0.03 (0.06) | 0.12 (0.13) | 0.84 (0.14) | 0.15 (0.13) |
| Physical | 5 | 0.72 (0.21) | 0.08 (0.12) | 0.00 (0.02) | 0.19 (0.17) | 0.80 (0.17) | 0.20 (0.17) |
| Information | 4 | 0.68 (0.22) | 0.13 (0.19) | 0.13 (0.18) | 0.04 (0.10) | 0.81 (0.18) | 0.19 (0.18) |
| Average domain ($SD$) performance[c] | | 0.69 (0.11) | 0.13 (0.09) | 0.06 (0.06) | 0.12 (0.09) | 0.82 (0.10) | 0.17 (0.11) |

CI = Critical items.

*Note.* Due to rounding, row percentages may not sum to 100.

[a] Refers to history taking, physical examination, and information sharing item content domains as described in Chapter 3.

[b] Average number of items by domain.

[c] Average proportion of items by domain.

Striking differences in agreement did exist, however, between domains. On average, students over-reported 13% of information sharing critical action items, much more than in the other two domains; information sharing items were rarely under-reported (4% of information sharing items). On the other hand, physical examination critical action items on average were never over-reported (0% of physical examination items), but were on average under-reported considerably (19% of physical examination items).

*Differences in agreement by content domain.* Results of a repeated measures ANOVA indicated no significant effect of domain on the agreement between medical student and expert rater,

$F(2, 136) = 0.98$, $p = .379$.

*Differences in over-reporting by content domains.* A doubly multivariate repeated measures ANOVA did reveal, however, a statistically significant effect of domain on levels of student and expert disagreement when examining both over-reporting and under-reporting. There was a statistically significant effect of domain on the proportion of critical action items over-reported,

$F(1.18, 87.24) = 30.86$, $p < .01$. [4] Pairwise comparisons revealed a significant difference in over-reporting between history and physical examination domain items, $p = .002$, and between history and information sharing domain items, $p < .01$, as well as between physical examination and information sharing domain items, $p < .01$. Students over-reported information sharing significantly more than both history and physical examination critical action

[4]Mauchly's test indicated that the assumption of sphericity for proportion over-reported had been violated, $\chi^2(2) = 87.02$, $p < .01$. Degrees of freedom were therefore corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = 0.59$).

items; in fact, students over-reported physical examination items rarely.

*Differences in under-reporting by content domain.* There was a statistically significant effect of domain on the proportion of critical action items under-reported, $F(1.82, 134.97) = 24.57$, $p < .01$. [5] Pairwise comparisons revealed a significant difference in under-reporting between history and physical examination domain items, $p = .022$, and between history and information sharing domain items, $p < .01$, as well as between physical examination and information sharing domain items, $p < .01$. Students under-reported physical examination significantly more than history and information sharing critical action items. These results suggested that while no domain displayed significantly more overall agreement between student and expert, there were significant differences between the domains in proportion of critical action items over- and under-reported.

### 4.3.2 Agreement between medical student and expert rater by features of examination

Parallel analyses revealed no significant differences in agreement, in over-reporting, or in under-reporting between the medical student and the expert rater by week of the examination (week 1, week 2, or week 3), the timing of the examination (morning or afternoon), or the order of the station (first, second, or third).

An increase in instances of over-reporting from the morning to afternoon session or from week to week of the examination, might have indicated student cheating as, in theory, stu-

[5] Mauchly's test indicated that the assumption of sphericity for proportion under-reported had also been violated, $\chi^2(2) = 9.48$, $p = .009$. Degrees of freedom were therefore corrected using Huynh-Feldt estimates of sphericity ($\epsilon = 0.91$).

dents who obtained information prior to completing the examination from fellow students may have inadvertently reported information that they did not learn in the encounter. Findings from this study provided no evidence of improvement in medical student performance over the course of the examination period; however, one unfortunate possibility is that past medical students have shared information about the stations with all students. Stations do oftentimes repeat year after year, meaning the entire population of medical students may already "know" some of the stations that appear on the examination from older classmates, though it is unlikely that students know the exact items that appear on the behavioral checklist as this information is strictly guarded.

### 4.3.3 Agreement between medical student and expert rater by characteristics of the medical student

Medical student characteristics, specifically social skills–as measured by Patient-Physician Information (PPI, see Chapter 3) as rated by the SP–and gender were also of interest. The relationship between the level of overall medical student agreement with the expert rater and PPI scores was small, negative, and not significant, $r(67) = -.11$, $p = .376$. There was no association between level of student social abilities, as measured by PPI, and student ability to report the encounter truthfully.

Results of an independent samples t-test revealed no significant difference in medical student agreement with the expert rater, nor in over-reporting, or in under-reporting by medical student gender. Additional analyses revealed no significant difference between male and female medical students in appropriately reporting, over-, and under-reporting informa-

tion from male and female SP encounters. Regardless of SP gender, male and female medical students documented with similar levels of agreement (with the expert rater) the details of the medical encounter.

### 4.3.4 Relationship between medical student performance and correct report

Surprisingly, competence, or performance of critical actions, based on expert observation, had no relationship to correct report of those critical actions, $r(69) = -.05$, $p = .708$. More competent students (again, based on expert scoring of performance), did not demonstrate a higher level of correctly reported information in their patient notes. Likewise, some poor performing students were also quite capable of correctly reporting their performance, while others were not. This may indicate that performance and report of performance are two distinct, and unrelated skills, which should not be conflated.

### 4.3.5 Summary of agreement between medical students and expert rater

In summary, there were significant differences in documentation of the clinical encounter between the medical students and the expert rater. Content of information reported, specifically station and content domain, did account for some of the differences in level of agreement between medical student and expert rater, whereas features of the examination and characteristics of the medical student did not. Levels of agreement between the students and the expert rater were significantly higher in station 2 (cough) than in station 1 (lower back

pain) and station 3 (heavy menses). While level of overall agreement between the medical student and expert did not differ significantly by domain, there were significant differences in level of disagreement between medical student and expert by domain, both in over-reporting and under-reporting. Disagreement over history and physical examination domain items was generally due to students failing to document behaviors the expert saw them perform in the encounter (under-reporting), whereas disagreement over information sharing items was generally due to students documenting behaviors the expert did not see them perform in the encounter (over-reporting). Station 3 displayed significantly higher levels of over-reporting, specifically of information sharing items. Station 1–the most difficult station for students and the station with the lowest overall agreement–displayed significantly higher levels of under-reporting among students.

These results indicated that agreement between medical student and expert rater was lowest in stations 1 and 3. Whereas students struggled in station 1 to report all critical information in their self-reports, they had difficulty sharing with the patient all critical information during the encounter in station 3. It is unclear why students found the greatest difficulty reporting information in station 1; perhaps, of all three stations, this station, which involved the interview of a patient experiencing severe back pain, was the most difficult for students to distinguish information that was critical from information that was not, leading to significantly higher levels of under-reporting and lower levels of agreement. It is plausible that students simply could not make sense of the patient's pain, or even lacked skills related specifically to the diagnosis and treatment of pain, and therefore did not know what to report.

Across domains, student levels of agreement with the expert rater were no different

overall; however, there were significantly higher levels of students under-reporting physical examination critical action items that they had performed. This may stem from medical students being unable to correctly interpret patient response to maneuvers during the physical examination. If a medical student does not note a positive finding from the physical examination (albeit incorrectly), they may be less likely to report that aspect of the examination in their self-report. Another plausible explanation is the misuse of special equipment provided during the encounter to enhance findings of the physical examination (See Table 4.2). In station 2, with the patient complaining of a troublesome cough, some students did not correctly report a physical examination maneuver that involved the use of special equipment (provided only for that encounter) that they had, in fact, performed.

Medical students also significantly over-reported more information sharing critical action items than history and physical examination items, particularly in station 3, with the patient complaining of a heavy menses. This indicated that medical students oftentimes failed to discuss with the patient in the encounter a possible diagnosis or important next steps but did note this information in the self-report. First and foremost, it is important to understand that students were instructed to provide a diagnosis to the patient in the encounter and to report in their note what that diagnosis was. One simple explanation for why students did over-report information sharing items is that they simply failed to follow instructions, though perhaps the reason behind this is more complex. While over-reporting could indicate falsification of the patient record (i.e., lying), it is possible that students were simply not as experienced at sharing information about upcoming diagnostic tests and potential diagnoses with the patient, a behavior that is likely left not to the medical student but to the attending physician in the clinic. Students probably have more experience sharing

their thoughts about diagnosis in the patient note than with an actual patient. Therefore, perhaps this preponderance of over-report simply reflected a very real phenomenon among students. Students had difficulty following instructions because experience has told them to behave otherwise. Another possibility is that students struggled to process in the moment the patient's history and physical examination findings to arrive at a plausible diagnosis and plan of treatment and so neglected to share with the patient a specific possible diagnosis, perhaps offering instead a vague diagnosis of symptoms in the encounter. By the time the student composed the patient note, they had had some opportunity to reflect and provide a specific diagnosis. Still another possibility is that students felt uncomfortable giving patients bad news, like the likelihood of a life-threatening illness. Rather than inform the patient that, for instance, cancer or HIV/AIDS, is a possibility, the student perhaps placed this information in the self-report to communicate critical information not with the patient but with fellow physicians, or in the case of this examination, the faculty evaluator. In station 3, the station with the highest level of over-reporting of information sharing items, discomfort delivering bad news to a woman complaining of heavy menstrual bleeding could account for the high level of over-reporting. This explanation seems less likely, though, as students were also required to deliver bad news in station 2 to the patient with an unusually bad cough and apparently had little difficulty doing so (see Table 4.4, station 2, item 4). These reasons could all explain why students failed to follow instructions during the examination and report on the potential diagnoses they had shared with the patient in the clinical encounter.

Remarkably, other contextual factors–like the date and time of the examination and specific characteristics of the medical student–did not explain a significant amount of variance in level of agreement between medical students and the expert. This affirms firstly that

students are likely not cheating; increased over-reporting, for instance, over the course of the examination period could indicate cheating. Secondly, these results suggest a lack of bias based on medical student characteristics like gender and social ability. Those students with higher social ability, as indicated by higher PPI scores, do not necessarily report any more or less correctly than those students with lower social ability. Likewise female medical students did not report more correctly than male medical students their encounter with the female patient in station 3 who was complaining of a heavy menses. These results only confirm the importance of examining other potential causes of incorrect report, beyond cheating, in order to improve medical student self-report.

These results are important for two reasons. Firstly, based on these findings, use of medical student self-report to score clinical competence in its present form on the CPX is not advisable. Too often, students under-reported or over-reported items, leading to a dramatic difference in the overall number of students who passed the examination based on scoring of the self-report. Secondly, even if not using the self-report to score performance on the examination, these results underscore the need for improved training, practice, and experience with the patient note as part of medical training. The lack of relationship between actual student performance, as observed by the expert, and medical student ability to truthfully capture the clinical encounter suggests that performance and report of performance are two separate skills, both important to clinical competence. Students clearly struggled to appropriately report information in the note, and considering the patient note is a key component of any physician's practice, it is imperative that students develop and hone this skill.

## 4.4 SP agreement with the expert rater

In this study of medical student self-report, it is also equally important to consider the agreement in report of medical student performance between SP and expert rater, as SP ratings are currently used by the institution, as well as at many other institutions, to furnish scores of medical student clinical competence. Examination of SP-expert agreement and disagreement provides a benchmark by which we can judge medical student self-report. With only a few exceptions, SP agreement with the expert rater items performed by the medical student was high. On average, SPs were in agreement with the expert rater on 92% of items, as seen in Table 4.7. Though there was little disagreement between SP and expert, over-reporting was more common than under-reporting (5.7% versus 2.3%, respectively, see Table 4.7).

Table 4.7: Agreement between SP and expert rater by item

| | | | | No. of SPs (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | In Agreement | | In Disagreement | | | | |
| Station | Item | Content domain[a] | $N$ | Observed by expert, reported by SP | Not observed by expert, not reported by SP | Not observed by expert, reported by SP (over-report) | Observed by expert, not reported by SP (under-report) | Total in agreement | Total in disagreement | Test statistic[b] |
| 1 | 1 | History | 72 | 71 (99) | 1 (1) | 0 (0) | 0 (0) | 72 (100) | 0 (0) | 0.00 |
| 1 | 2 | History | 72 | 56 (78) | 6 (8) | 9 (13) | 1 (1) | 62 (86) | 10 (14) | 4.90 |
| 1 | 3 | History | 72 | 24 (33) | 42 (58) | 6 (8) | 2 (3) | 64 (89) | 8 (11) | 0.17 |
| 1 | 4 | History | 72 | 51 (71) | 5 (7) | 15 (21) | 1 (1) | 56 (78) | 16 (22) | 9.60* |
| 1 | 5 | Physical | 72 | 39 (54) | 16 (22) | 0 (0) | 17 (24) | 55 (76) | 17 (24) | 16.06* |
| 1 | 6 | Physical | 72 | 64 (89) | 8 (11) | 0 (0) | 0 (0) | 72 (100) | 0 (0) | – |
| 1 | 7 | Information | 72 | 70 (97) | 0 (0) | 1 (1) | 1 (1) | 70 (97) | 2 (3) | 0.50 |
| 2 | 1 | History | 74 | 74 (100) | 0 (0) | 0 (0) | 0 (0) | 74 (100) | 0 (0) | – |
| 2 | 2 | History | 74 | 71 (96) | 0 (0) | 3 (4) | 0 (0) | 71 (96) | 3 (4) | – |
| 2 | 3 | Physical | 74 | 74 (100) | 0 (0) | 0 (0) | 0 (0) | 74 (100) | 0 (0) | – |
| 2 | 4 | Information | 74 | 67 (91) | 4 (5) | 3 (4) | 0 (0) | 71 (96) | 3 (4) | 1.33 |
| 3 | 1 | History | 75 | 75 (100) | 0 (0) | 0 (0) | 0 (0) | 75 (100) | 0 (0) | – |
| 3 | 2 | History | 75 | 62 (83) | 5 (7) | 3 (4) | 5 (7) | 67 (89) | 8 (11) | 0.13 |
| 3 | 3 | History | 75 | 70 (93) | 2 (3) | 3 (4) | 0 (0) | 72 (96) | 3 (4) | 1.33 |
| 3 | 4 | Physical | 75 | 72 (96) | 3 (4) | 0 (0) | 0 (0) | 75 (100) | 0 (0) | – |
| 3 | 5 | Physical | 75 | 73 (97) | 2 (3) | 0 (0) | 0 (0) | 75 (100) | 0 (0) | – |
| 3 | 6 | Information | 75 | 36 (48) | 8 (11) | 29 (38) | 2 (3) | 44 (59) | 31 (41) | 21.81* |
| 3 | 7 | Information | 75 | 44 (59) | 25 (33) | 5 (7) | 1 (1) | 69 (92) | 6 (8) | 1.50 |
| Average total (%) performance[c] | | | | 14.8 (82.1) | 1.8 (9.9) | 1.0 (5.7) | 0.4 (2.3) | 16.6 (92.0) | 1.4 (8.0) | |

*$p < .003$.

*Note.* Due to rounding, row percentages may not sum to 100.

[a] Refers to history taking, physical examination, and information sharing item content domains as described in Chapter 3.

[b] McNemar's test statistic calculated using the continuity correction to determine if data exhibit marginal homogeneity. The test statistic was compared to the $\chi^2$ distribution with 1 degree of freedom.

[c] Average number of 18 items.

In general, in instances of disagreement, SPs tended to over-report information about the medical students, with the exception of station 1, item 5 (discussed below). It is important to note, trainers instructed SPs to give medical students credit for critical action items when in doubt. If the SP could not recall when completing the checklist whether or not a medical student had in fact performed a specific behavior, the SP was to assign the student credit. This partially explains instances of over-reporting among SPs.

Items with the highest level of disagreement between SP and expert rater were: station 1, item 4; station 1, item 5; and station 3, item 6. Based on a series of McNemar's tests comparing the dependent proportions of students who SPs documented performing a critical action item to those students who were observed performing a critical action item by the expert rater, these item-level differences between the SP and the expert rater were determined significant at the $p < .003$ level ($p$-value adjusted using Bonferroni correction). As discussed below, these items all exhibited poor clarity, which may explain the high disagreement.

Comparison of SP and expert rater documentation revealed substantial over-reporting for station 1, item 4 (21% of students, history, see Table 4.7) and, even more so, for station 3, item 6 (38% of students, information sharing, see Table 4.7). For these items, SPs erroneously gave students credit for obtaining patient history and for sharing information in the encounter when students had, in fact, not asked the relevant question or shared the relevant information. Both of these items required the SP to document more than one behavior. For station 1, item 4, the SP was prompted to indicate whether or not the medical student performed multiple behaviors (e.g., 'asked me to describe A, B, and C'). In station 3, item 6, the SP was prompted to indicate whether or not the medical student had shared at least two pieces of information from a longer list of information (e.g., 'told me symptoms indicated a

diagnosis of TWO of the following: A, B, C, D, E'). The wording of these items may have confused SPs. In the case of station 1, item 4, what if a student asked for a description of A and C, but not B? Perhaps the SP could not remember whether the student had mentioned C, but could clearly recall A and B and therefore assigned the student credit erroneously for the behavior. The wording of station 3, item 6 may have also confused SPs. Perhaps they recalled one of the listed diagnoses but could not clearly recall a second and so assigned the student credit for performing the item.

In contrast, SPs under-reported nearly 25% (17 of 72) of student performance in station 1, item 5, a physical examination item (e.g., 'performed one maneuver OR another maneuver on BOTH sides of the body'). Though the student performed one of these the maneuver, the SP failed to give him or her credit for the item. Again, the wording of this item, which asked for three pieces of information–whether the maneuver had been performed, whether a different but equal maneuver had been performed, and if either was performed on both sides of the body–may have created difficulties for the SP in scoring medical student performance.

Clarity can easily be addressed by assigning one behavior to each performance item. For instance, station 1, item 4 (see Table 4.7) can be broken into three separate items, so that SPs report not that a student 'asked me to describe A, B, and C,' but rather 'asked me to describe A,' 'asked me to describe B,' and 'asked me to describe C.' Alternatively, faculty may determine that one or two of these items are more important than the others; in that case, the behavioral checklist should include only that item (e.g., 'asked me to describe B.'). Station 1, item 5 also was complicated in its wording, asking whether or not the student 'performed one maneuver OR another maneuver on BOTH sides of the body.' Again, this item could be broken down into minimum two items: 1) 'performed maneuver on BOTH

sides of the body;' and 2) 'performed other maneuver on BOTH sides of the body.' Later, evaluators compiling scores can easily award students credit for performing one or the other of these items.

Item clarity, however, cannot explain all of SP difficulty correctly reporting student behavior in these items. Though the aforementioned items did exhibit poor clarity, other items on the behavioral checklist also lacked clarity but were correctly reported in high levels by the SP. This includes station 1, items 6 and 7 and station 3, item 4. SPs were in 100% agreement with the expert on report of performance for station 1, item 6. Like station 1, item 5, this item asked whether or not students had performed at least one of two maneuvers on both sides of the body. Interestingly, of the two maneuvers that were acceptable for credit, students performed mainly one and not the other, so SPs never had to determine if students had performed one of two maneuvers. Item station 1, item 7, an information sharing item, asked whether or not the student had told the patient he might have A or B or C. SPs had very high agreement with the expert (97%) despite the wording. Unlike station 3, item 6, however, this item asked only for one of three correct diagnoses, not two of five. Finally, station 3, item 4 asked whether or not the student had performed certain aspects of a physical examination (e.g., 'performed maneuver A, maneuver B, and maneuver C'). Perhaps the examination was so rote for students that once students initiated the general physical examination required in station 3 (and most all students did), they performed a complete physical examination, including all maneuvers listed in the item. Therefore, SPs, even in instances of doubt, were more likely to correctly award credit because students performed a complete physical. This is in contrast to station 1, item 5, which also asked if a student had performed one of two maneuvers; however, the maneuvers in station 1 were not part of a

larger general physical examination. En sum, careful attention must be paid to the wording of each individual item on the behavioral checklist with consideration given to how students are likely to perform an item in a given station. Items with low agreement between SP and expert may be the result of not clarity alone, but of other mitigating factors.

### 4.4.1 Agreement between SP and expert rater by content and context of information

Table 4.8 displays the agreement of SP with the expert rater in documenting the clinical encounter by station. On average, SP agreement with the expert by station was extremely high (93% of items in each station). Again, over- reporting was more problematic for SPs than under-reporting (5% versus 2%, respectively, of items on average per station).

Table 4.8: Level of agreement between SP and expert rater by station

| | | | *M* (*SD*) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | In Agreement | | In Disagreement | | | |
| Station | CI | *N* | Observed by expert, reported by SP | Not observed by expert, not reported by SP | Not observed by expert, reported by SP (over-report) | Observed by expert, not reported by SP (under-report ) | Total in agreement | Total in disagreement |
| | | | No. of items | | | | | |
| 1 | 7 | 72 | 5.2 (1.1) | 1.1 (0.9) | 0.4 (0.6) | 0.3 (0.5) | 6.3 (0.8) | 0.7 (0.8) |
| 2 | 4 | 74 | 3.9 (0.3) | 0.1 (0.2) | 0.1 (0.3) | 0.0 (0.0) | 3.9 (0.3) | 0.1 (0.3) |
| 3 | 7 | 75 | 5.8 (1.0) | 0.6 (0.7) | 0.5 (0.6) | 0.1 (0.4) | 6.4 (0.7) | 0.6 (0.7) |
| Average station (*SD*) performance[a] | | | 4.9 (0.5) | 0.6 (0.4) | 0.3 (0.3) | 0.1 (0.2) | 5.5 (0.3) | 0.5 (0.3) |
| | | | Proportion of items | | | | | |
| 1 | 7 | 72 | 0.75 (0.16) | 0.15 (0.12) | 0.06 (0.08) | 0.04 (0.07) | 0.90 (0.12) | 0.10 (0.12) |
| 2 | 4 | 74 | 0.97 (0.09) | 0.01 (0.06) | 0.02 (0.07) | 0.00 (0.00) | 0.98 (0.07) | 0.02 (0.07) |
| 3 | 7 | 75 | 0.82 (0.15) | 0.09 (0.10) | 0.08 (0.09) | 0.02 (0.06) | 0.90 (0.10) | 0.09 (0.10) |
| Average station (*SD*) performance[b] | | | 0.84 (0.08) | 0.09 (0.06) | 0.05 (0.04) | 0.02 (0.03) | 0.93 (0.05) | 0.07 (0.05) |

CI = Critical items.

*Note.* Due to rounding, row percentages may not sum to 100.

[a] Average number of items by station.

[b] Average proportion of items by station.

In station 2, the SP and expert rater had very high agreement, $M = 0.98$, $SD = 0.07$. Agreement between SP and expert rater in stations 1 and 3 was somewhat lower, $M = 0.90$, $SD = 0.12$ and $M = 0.90$, $SD = 0.10$, respectively, but still high. Analyses first addressed differences by station in levels of agreement between SP and expert using repeated measures ANOVA, [6] then differences in disagreement, specifically under- and over-reporting, using doubly multivariate repeated measures ANOVA. Parallel analyses examined differences in levels of agreement and levels of disagreement by domain.

*Differences in agreement by station.* A repeated measures ANOVA was performed to examine differences in agreement between the SP and expert rater by station. There was a statistically significant effect of station on the proportion of critical action items reported appropriately by the SP, $F(1.74,\ 122.11) = 13.04$, $p < .01$. [7] Pairwise comparisons revealed a statistically significant difference in the agreement between SP and expert rater documentation of station 1 and station 2, $p < .01$, and between station 3 and station 2, $p < .01$.

*Differences in over-reporting by station.* The type of SP disagreement with the expert rater also differed significantly by station. A doubly multivariate repeated measures ANOVA was employed to examine differences in disagreement between the SP and expert rater. The results indicated a statistically significant effect of station on the proportion of critical action items over-reported by the SP, $F(2,\ 140) = 8.70$, $p < .01$. Pairwise comparisons revealed small significant differences in SP over-reporting between station 1 and station 2, $p = .048$,

[6] Though measures of agreement between the SP and the expert rater were not normally distributed, use of non-parametric statistical techniques produced results similar to those reported here.

[7] Mauchly's test indicated the assumption of sphericity had been violated, $\chi^2(2) = 13.04$, $p = .001$. Degrees of freedom were therefore corrected using Huynh-Feldt estimates of sphericity ($\epsilon = 0.87$).

and between station 3 and station 2, $p < .01$.

*Differences in under-reporting by station.* The results indicated a statistically significant effect of station on the proportion of critical action items SPs under-reported, $F(1.57, 109.62) = 11.76$, $p < .01$. [8] Pairwise comparisons revealed a small significant difference in under-reporting between station 1 and station 2, $p < .01$. It is important to note that though there were significant differences in agreement between SPs and expert rater between stations, these differences were relatively small.

Parallel analyses examining the effect of domain on proportion of critical action items between SP and expert rater in agreement, over-, and under-reported were also conducted. Table 4.9 displays the agreement between SP and expert rater by domain.

[8] Mauchly's test indicated the assumption of sphericity for SP proportion under-reported had been violated, $\chi^2(2) = 24.63$, $p < .01$. Degrees of freedom were therefore corrected using Huynh-Feldt estimates of sphericity ($\epsilon = 0.78$).

Table 4.9: Level of agreement between SP ($N = 71$) and expert rater by domain

| | | Mean (SD) | | | | | |
|---|---|---|---|---|---|---|---|
| | | In Agreement | | In Disagreement | | | |
| Content domain[a] | CI | Observed by expert, reported by SP | Not observed by expert, not reported by SP | Not observed by expert, documented by SP (over-report) | Observed by by expert, not reported by SP (under-report) | Total in agreement | Total in disagreement |
| | | No. of Items | | | | | |
| History | 9 | 7.5 (0.9) | 0.9 (0.8) | 0.5 (0.6) | 0.1 (0.3) | 8.4 (0.7) | 0.6 (0.7) |
| Physical | 5 | 4.4 (0.7) | 0.4 (0.6) | 0.0 (0.0) | 0.2 (0.4) | 4.8 (0.4) | 0.2 (0.4) |
| Information | 4 | 2.9 (1.0) | 0.5 (0.7) | 0.5 (0.6) | 0.1 (0.2) | 3.4 (0.6) | 0.6 (0.6) |
| Average domain ($SD$) performance[b] | | 4.9 (0.5) | 0.6 (0.4) | 0.3 (0.3) | 0.1 (0.2) | 5.5 (0.3) | 0.5 (0.3) |
| | | Proportion of Items | | | | | |
| History | 9 | 0.84 (0.10) | 0.10 (0.08) | 0.05 (0.07) | 0.01 (0.04) | 0.93 (0.08) | 0.07 (0.08) |
| Physical | 5 | 0.87 (0.13) | 0.08 (0.12) | 0.00 (0.00) | 0.05 (0.08) | 0.95 (0.08) | 0.05 (0.08) |
| Information | 4 | 0.72 (0.24) | 0.13 (0.18) | 0.13 (0.15) | 0.02 (0.06) | 0.85 (0.16) | 0.15 (0.16) |
| Average domain ($SD$) performance[c] | | 0.81 (0.10) | 0.10 (0.08) | 0.06 (0.05) | 0.02 (0.04) | 0.91 (0.06) | 0.09 (0.06) |

CI = Critical items.

*Note.* Due to rounding, row percentages may not sum to 100.

[a] Refers to history taking, physical examination, and information sharing item content domains as described in Chapter 3.

[b] Average number of items by domain.

[c] Average proportion of items by domain.

*Differences in agreement by content domain.* A repeated measures ANOVA was employed to examine the effect of domain on level of agreement between SP and expert rater. Results indicated a significant effect of domain on SP-expert rater agreement, $F(1.44, 100.80) = 15.01$, $p < .01$. [9] Pairwise comparisons revealed statistically significant differences in agreement between SP and expert rater between information sharing and history items, $p = .001$, and between information sharing and physical examination domain items, $p < .01$. SPs were significantly less in agreement with the expert rater on information sharing items–items pertaining to student discussion of potential diagnosis and treatment–than on other domain items.

*Differences in over-reporting by content domain.* A doubly multivariate repeated measures ANOVA also revealed a statistically significant effect of domain on SP over- and under-reporting. There was a statistically significant main effect of domain on the proportion of critical action items SPs over-reported, $F(1.24, 87.08) = 34.25$, $p < .01$. [10] Pairwise comparisons revealed a significant difference in SP over-reporting between history and physical examination domain items, $p < .01$, and between history and information sharing domain items, $p = .001$, as well as between physical examination and information sharing domain items, $p < .01$. SPs over-reported information sharing items information sharing items significantly more than history and physical examination items. In fact, SPs did not over-report a single physical examination item.

---

[9] Mauchly's test indicated the assumption of sphericity had been violated, $\chi^2(2) = 33.99$, $p < .01$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = 0.72$).

[10] Mauchly's test indicated that the assumption of sphericity for the proportion SP over-reported items had been violated, $\chi^2(2) = 64.56$, $p < .01$. Degrees of freedom were therefore corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = 0.62$).

*Differences in under-reporting by content domain.* There was a statistically significant main effect of domain on the proportion of critical action items under-reported, $F(1.26, 110.40) = 6.10$, $p = .006$. [11] Pairwise comparisons revealed a significant difference in under-reporting physical examination and history domain items, $p = .018$, and between physical examination and information sharing domain items, $p = .037$. SPs under-reported physical examination domain items significantly more than history and information sharing critical action items. Note that the largest item under-reported by SPs–station 1, item 5– was a physical examination item.

In summary, while SPs agreed highly with the expert rater, on average, there were small, but significant, differences in their level of agreement by domain. SPs disagreed with the expert rater more often on items from the information sharing domain than from the physical and history domains. There are a couple of potential explanations for this discrepancy. Firstly, it is important to recall that SPs were trained in instances of doubt to award students credit for performing a critical action item. Perhaps SPs could recall that the student had discussed, vaguely, thoughts about a potential diagnosis, but could not recall whether or not they had provided a specific diagnosis, let alone an accurate diagnosis, and in doubt awarded students credit for the item. Alternatively, considering the fact that information sharing generally occurred towards the end of the encounter, poor agreement in this domain may indicate SP fatigue within the encounter. By the end of the encounter, perhaps the SP could not precisely recall medical student performance on these items, leading to over-report

[11] Mauchly's test indicated that the assumption of sphericity for proportion under-reported items had also been violated, $\chi^2(2) = 23.77$, $p < .01$. Degrees of freedom were therefore corrected using Huynh-Feldt estimates of sphericity ($\epsilon = 0.79$).

(giving students the benefit of the doubt). SP under-reporting proved more of a problem for physical examination items, though mean differences between physical examination items and the other two domains were quite small. This could indicate training deficiencies in the SPs, who are unable to correctly recognize physical examination behaviors, and therefore neglect to award students credit for performing these items. Finally, it is important to note that those items with poor agreement between SP and expert rater also exhibited poor clarity in their wording.

### 4.4.2 Agreement between SP and expert rater by features of the examination

Parallel analyses revealed no significant differences in agreement between the SP and the expert rater nor in over- and under-reporting by week of the examination (week 1, week 2, or week 3), the timing of the examination (morning or afternoon), or station order (first, second, or third).

### 4.4.3 Agreement between SP and expert rater by characteristics of the medical student

The relationship between SP correct report and medical student characteristics like social skills (as measured by PPI) and gender were also examined. The relation between SP-expert agreement over all critical items and PPI scores (as rated the SPs) was small, negative, and not significant, $r(69) = -.05$, $p = .706$. There was no association between student social competence and the level of agreement between SP and expert. SPs did not over-

report performance of items by students with higher social skills nor under-report items by students with lower social skills.

Results of an independent samples t-test revealed no significant difference in SP agreement with the expert rater, nor in over-reporting, nor in under-reporting by medical student gender. Additional analyses revealed no significant difference between male and female SPs in agreement with the expert rater, nor in over-, and under-reporting information from encounters with male and female medical students. Regardless of medical student gender, male and female SPs documented with similar levels of agreement (with the expert rater) the details of the medical encounter.

### 4.4.4 Summary of agreement between SP and expert rater

In summary, as with the medical students, there were instances of disagreement between the SP and expert rater in reporting the clinical encounter, though SPs were more in agreement with the expert rater than medical students on individual items. Significant differences in agreement were found between stations and domains, but not between features of the examination and characteristics of the medical student. Both medical students and SPs found station 2, involving a man complaining of a persistent cough, the easiest to document correctly, suggesting perhaps that an inherent quality of that station (e.g, chronic illness, the critical action items themselves, SP training, etc., see additional station differences in Table 4.3) promoted a higher level of report of information. Finally, it is possible that changes to the wording of some critical items, which lacked clarity, might improve SP ability to correctly score student performance for those items.

Based on these findings, although the SPs at the item-level were more correct on average in their report of student performance than medical students, incorrect report still identified a substantial number of students as passing the examination who, based on standards determined by faculty, should have failed the examination. SP documentation must be improved. Firstly, educators should re-examine each station protocol, or the description of the patient character provided to each SP and used as the basis for the station, specifically focusing on any material related to the critical action items to ensure that this information is clearly detailed. Secondly, educators need to revisit the behavioral checklist to simplify, pilot, and implement changes to those items with potential clarity issues. SPs were on average more correct in their report of student behavior; however, the dramatic difference in the number of students who passed the examination based on SP documentation as compared to expert observation is cause for concern and must be addressed.

## 4.5 Chapter summary

Significant differences existed between the different encounter participants, both between the medical student and between the SP and the expert rater, in the recording of behaviors. In medical education, currently many rely on the SP to report student behaviors in a performance examination encounter. In some circumstances, institutions may use faculty or other expert raters to score student performance, though this is often not feasible given the added associated costs and faculty time constraints. Use of medical student self-report to evaluate competence has gained some traction thanks to the United States Medical Licensing Examination (USMLE), though this method is not currently in widespread use among

medical schools administering high-stakes clinical competence examinations like the CPX. Results presented here indicated that medical students have difficulty correctly capturing what happened in an encounter, meaning the use of self-report to assess professional competence has its limitations. Though discrepancies did exist between the SP and the expert rater, there were even more so between the medical student and the expert rater.

*Agreement between medical student and expert rater.* Medical student performance of critical action items, based on observation by the expert rater, was quite high. Disagreement between the medical student and the expert rater was, unfortunately, also high, resulting in a substantial difference in the number of students who passed the examination based on the medical student self-report and based on expert observation. This study explored several contextual factors that may account for discrepancies between medical student and expert observer in the report of an encounter, including content and context of information, features of the encounter, and characteristics of the medical student under assessment. Content and context of information reported–specifically station and critical action item domain– alone accounted for significant differences in levels of agreement and disagreement with the expert rater, whereas features of the encounter–day of the examination, time of the examination, order of the encounter–and characteristics of the medical student–gender, level of social ability–did not.

Medical students demonstrated high levels of agreement with the expert rater in station 2–the station involving a man with a bad cough–indicating perhaps that some feature of this station promoted greater recollection of critical details. For the medical students, increased levels of agreement could indicate not only that students were more correct in their recollection of behaviors, but also that they were more readily able in station 2 to distinguish critical

details to document in their self-reports. Possible explanations include inherent features of the station, like its portrayal of a chronic condition or the use of special equipment to enhance physical examination findings, as well as measurement issues, like the relative quantity of critical information. Of the three stations, station 2 had the least amount of critical action items, nearly half the amount of each of the other two stations. Perhaps there is a certain threshold of information that students can correctly recall in any given encounter. Whatever the cause, different contexts within a behavioral assessment, like station 2 as opposed to stations 1 with the man complaining of back pain and 3 with the woman complaining of an unusually heavy menses, may yield different levels of correct student self-report, which is important to account for in any evaluation of competence.

With regard to the kind of information collected about professional behaviors, medical students were prone to under-reporting critical action items, particularly those related to history and physical examination. Despite instructions to share with the patient information about diagnosis and then to indicate that information in the patient note, medical students significantly over-reported information sharing items, meaning they neglected to discuss with the patient during the encounter specific diagnoses and next steps yet reported this information in the patient note. Possible explanations include student inexperience with sharing information with the patient directly or student discomfort in delivering bad news. It is even possible that students felt they had shared specific information about a potential diagnosis and next steps, but they were, in fact, too general in their description of the illness to the patient and more specific, or technical, in their patient note, which is addressed to other physicians. Whatever the cause, in addition to context (i.e., station), the content of the item of information reported also mattered to its correct report. Overall, medical students

struggled considerably, particularly when compared to SP report.

*Agreement between SP and expert rater.* Because SP report currently serves as the basis for student performance scores on the CPX, it was important to consider how student-expert agreement compared to SP-expert agreement. Surprisingly, SP agreement with the expert rater also revealed errors in SP report. Despite overall high levels of agreement with the expert rater, these errors resulted in a substantial number of students passing the examination who should have, in fact, failed. Like with medical students, content and context of information explained a significant (albeit small) proportion of disagreement in the report. Reassuringly, features of the encounter and characteristics of the medical student did not explain significant levels of agreement or disagreement, indicating that SPs maintained their ability to report over the course of the examination period and that they were not biased in their report of performance based on medical student gender or social skill level.

Like the medical students, SPs were more correct in their report of medical student behaviors in station 2 than in the other two stations. This may suggest, again, that some inherent feature of this station promotes correct report. Perhaps SPs, like medical students, can only correctly recall so much information, and since station 2 had the least number of critical action items, it may have just been easier for SPs to report performance in that station. Alternatively, discrepancies between the SP and the expert rater may indicate SP difficulty interpreting medical student behaviors in those stations in which students did not perform as well, particularly station 1 but also station 3. Student difficulty navigating station 1, for instance, as indicated by lower levels of item performance and difficulty correctly reporting performance, may have caused SPs to also struggle to correctly report performance. In other words, confusion on the part of students about how to interview and examine a

patient complaining of pain may have contributed to confusion on the part of SPs.

Though SPs enjoyed high levels of agreement with the expert rater, they did exhibit poor ability to correctly report information sharing items, over-reporting often that students had told them the potential diagnosis when they had, in fact, not. It is important to note that SPs were instructed to give students the benefit of the doubt, and therefore over-report, when having difficulty recalling a student's behavior. However, with the exception of information sharing items, there was little evidence of over-reporting among SPs. This either indicates that SP instruction to over-report was not necessarily the cause of over-report of information sharing items, since over-reporting was not widespread, or that SPs had greater difficulty recalling student performance of information sharing items. Considering information sharing occurs generally at the end of the interview, this may indicate a certain level of SP fatigue during an encounter, or that medical student information sharing was difficult for the SP to interpret. Finally, though some of these discrepancies between SP and expert could be explained by poor clarity in the wording of critical action items, other items also exhibited poor clarity and had high levels of SP-expert agreement, meaning poor clarity cannot be the sole cause of report error among SPs.

In the end, both SPs and medical students do not report performance correctly, calling into question the ability of either to capture true medical student performance using the critical action items approach. Medical students on average were less in agreement overall with the expert rater, leading to substantial differences in the number of students who failed the examination based on the medical student self-report. Though SPs were more in agreement on average with the expert rater, error in SP report lead to equally large number of students passing the examination who should have failed based on expert observation.

While SP error may be partially addressed by making changes to the behavioral checklist, medical student error in reporting of performance may require changes to the curriculum as well as changes to the actual examination. The lack of relationship between performance and correct report among medical students is troubling as it suggests that the actual performance of behaviors in a clinical encounter and the report of performance in the patient note are two distinct, unrelated skills. It calls into question the ability of self-report to truthfully capture student clinical competence; rather, self-report may reflect simply medical student ability to report performance in an encounter. Some students excel at both or neither, while others are better at only one.

# Chapter 5

# Discussion

This study explored the use of self-report to score performance and determine medical student competence, focusing on two major research questions. Firstly, what is the level of agreement between medical student self-report and expert rater documentation of a clinical encounter? Secondly, what is the level of agreement between a trained participant observer in the encounter, known as the standardized patient (SP), which is the most commonly used method for scoring student performance, and expert rater documentation of medical student performance? Additionally, this study investigated whether levels of agreement between both student and expert and between SP and expert depended on the context and content of information, features of the examination, or characteristics of the professional.

Use of performance evaluation among professionals has increased in popularity. Given the complexity of skills oftentimes required of professionals, proponents argue that these methods are better suited to situations when assessing knowledge is not enough to determine competence. Ensuring the reliability and validity of performance scores is vital to their accepted and ongoing use. Several methods of performance assessment attempt to cap-

ture professional behaviors. The self-report is one such method, wherein the professional is tasked with reflecting on practice and documenting behaviors, in the form of a questionnaire, attitudinal survey, directed prompts, free response, or other such strategies. This is in contrast to other methods that rely on participant evaluation, like SP ratings, and third-party observation of behaviors. In fact, some organizations like the National Board of Medical Examiners (NBME) are moving towards using the medical student self-report, or patient note, to score performance on the Step 2 Clinical Skills Licensing Examination. Though self-report has theoretical and practical advantages to assessing professional competence, it is also important to understand its limitations to ensure best practices and optimize its use. While the results of this study are directly pertinent to those in the medical profession, its findings can also be applied more broadly, offering guidance on issues related to the use of self-report in other professions such as education and teacher professional development.

The Clinical Performance Examination (CPX) is administered annually to all fourth-year medical students at the institution studied here. A high-stakes examination, the CPX is used to make decisions regarding students' clinical competence, or their ability to provide patient care. This study examined performance data from a stratified random sample of 75 fourth-year medical students who completed the examination in 2012. As part of this examination, students rotated through a series of 15-minute clinical encounters, called stations, to interview and examine a standardized patient (SP), an actor highly trained to portray a patient with a specific set of symptoms in a consistent and believable fashion. As part of the examination, students were instructed to: 1) interview the patient for pertinent information regarding the patient's history; 2) conduct a complete and thorough, focused physical examination; and 3) share with the patient important information regarding potential diagnoses

and next steps. Following the encounter, the SPs rated student performance using a behavioral checklist as well as rating student social ability, or their physician-patient interaction (PPI), while students completed a post-encounter activity. In 3 of the 8 post-encounter activities, students completed a self-report (called a patient note), detailing the history they obtained, the physical examination maneuvers they performed, and the diagnoses they shared with the patient.

Student clinical competence was scored based on performance of a set of behaviors in each station. The behavioral checklists for each station were comprised of, on average, 25 items spanning three content domains of the clinical encounter: history, physical examination, and information sharing. This study focused on student performance of a subset of those items –items deemed critical to patient care by an expert panel of faculty–which were used in creating a criterion-reference standard for the examination by faculty educators. This study examined student performance of these critical items as observed by an expert rater, as reported by the student in the patient note self-report, and as reported by the SP in each encounter.

While the three stations included in this study each involved patients seen in a medical clinic, there were some distinct differences between the three. Station 1 involved a 32 year old male patient complaining of lower back pain. Students were expected to perform a musculoskeletal examination. Based on the patient's history and physical examination findings, potential diagnoses did not include a life-threatening illness. The critical action items checklist for station 1 included 7 items. Station 2 involved a 50 year old male patient complaining of a chronic worsening cough. Students were expected to perform a cardiopulmonary examination using special equipment supplied to all students during the encounter to simulate

physical examination findings in the patient when there were, in fact, none. The student's list of potential diagnoses should have included some life-threatening illnesses. The critical action items checklist for station 2 was short in comparison, containing only 4 items. Station 3 involved a 44 year old female patient complaining of an unusually heavy menses. Students were expected to perform an abdominal examination. Students also received from the SP additional test results, key to correctly diagnosing the patient, if they mentioned this test to the SP. Diagnoses for this patient did include the possibility of life-threatening illness. The critical action items checklist for station 3 was 7 items long.

To evaluate medical student ability to correctly report their own performance, this study collected data about student performance on critical action items from three sources: the medical student self-report, the SP checklist, and expert rater observation of video recordings made of each patient encounter. Variables were then created based on the "match" between the student and the expert and between the SP and the expert, indicating where the student or SP had agreed with the expert rater, where they had over-reported, or documented a behavior that was not observed by the expert rater, and where they had under-reported, or failed to document a behavior that was, in fact, observed by the expert rater. This study also investigated whether or not contextual variables of interest could explain differences in level of agreement. These variables fell into three broad categories: 1) context and content of information, 2) features of the examination, and 3) characteristics of the medical student. More specifically these variables included: the station (station 1, 2, or 3), the item content domain (history, physical examination, information sharing), examination date (week 1, 2, or 3), examination time (morning or afternoon), station order (first, second, third), medical student gender, the interaction of medical student gender and SP gender, and student social

ability (as measured by PPI).

Results demonstrated that medical students both under-reported a considerable number of critical action items and over-reported a substantial number of critical action items. Of all contextual factors assessed in this study, only those related to context and content of information (i.e., station and content domain) explained a significant amount of variation in levels of agreement and/or disagreement between the medical student and the expert rater. Students correctly reported in high levels in station 2, over-reported significantly more in station 3, and under-reported significantly more in station 1. With regard to the content of information reported, students over-reported information sharing items significantly more and under-reported physical examination items significantly more than the other two domains.

It was also important to examine student agreement with the expert in relation to SP agreement with the expert, as SP ratings are most commonly used to construct performance scores and therefore serve as a good basis of comparison for student-expert agreement. Though medical student self-reports oftentimes incorrectly represented the clinical encounter, SP reports also, surprisingly, misrepresented student performance, when compared to expert rater observation. Educators wary of using medical student self-report to score performance must also appreciate that SP report of critical action items was not without its own flaws and was not necessarily a better alternative.

Scores based on medical student self-report did not agree with scores based on expert observation, resulting in a large number of students falsely identified as failing the performance examination based on their self-report. Considerable under-reporting by students resulted in an overall pass rate (18%) that was substantially less than that of the expert rater (42%).

Use of medical student self-report, therefore, to construct performance scores did have meaningful and, ultimately, negative consequences for the determination of competence.

The discussion that follows expands on these results, exploring specifically: 1) potential causes of medical student under- and over-report; 2) strategies to improve student ability to correctly self-report; and 3) areas of further study and research.

## 5.1 Causes of incorrect medical student self-report

*Differences in agreement in medical student self-report.* It is not immediately clear why medical students were prone to incorrectly report performance in their self-reports, though there are several plausible explanations. This discussion will focus on likely causes of poor agreement, specifically those related to curriculum and training, poor modeling by faculty, student performance in the individual stations, and the evaluation tools used in the assessment.

Students may lack the training necessary to successfully report on a clinical encounter. Results indicated no relationship between student performance and level of correct report, meaning high performing students did not necessarily report more correctly than their low performing counterparts. Students who demonstrated ability to perform critical action items did not necessarily demonstrate ability to correctly report those items. Performance and correct report, then, may constitute distinct skills, each requiring development in the young medical professional. Training and clinical experiences may have prepared students to perform appropriate behaviors, but not necessarily to report them.

Another, more troubling cause, of student inability to correctly report performance may

relate to poorly modeled self-report by interns, residents, and attending physicians. While students see physicians interact with patients, they may not often see them compose a patient note, as composing and writing a patient note is largely a mental activity; it is therefore not entirely an observable skill. Also possible is that physicians, if they do model this behavior, are themselves constructing poor quality patient notes that do not reflect the true encounter with the patient. After all, research shows that physicians oftentimes document details of the medical encounter incorrectly in the patient medical record (Berwick, 2002; West et al., 2002). Students may have either no model or a poor model of correct report of a patient encounter.

Correct report may also be tied to station difficulty, as the more difficult the station, the more difficulty students had in correctly reporting performance in their self-reports. Station 2, involving the man with a bad cough, was easiest for students, as indicated by expert observations of performance of critical action items in that station. In turn, students had significantly higher levels of agreement with the expert rater in reporting information from station 2. In contrast, station 1, which involved the man with lower back pain, was much more difficult for students, who were observed performing on average proportionately fewer items in that station and, in turn, correctly reported significantly fewer items in their self-reports, even for items they did, in fact, perform.

One possible explanation for such a relationship between station difficulty and student level of agreement with the expert rater may involve student training and their development of story schema, or student understanding of the world and, more specifically, illness. Effective schema, which are based on experience, greatly assist in the retelling of stories. If a particular patient narrative follows a student's schema of illness, known as an illness

script in medical education, then students may be better able to report back that story in the self-report. Though in their fourth year of medical training, student have had relatively few encounters (when compared to a practicing physician) with actual patients in a clinical setting. Rather, much of their training, particularly in the first two years, has focused on acquiring knowledge. As a result, in these clinical encounters, students are likely relying on knowledge and not experience to interview the patient. If a patient's presenting symptoms, therefore, do not align well with that knowledge (and limited experience), then the student's performance in the encounter may suffer and, subsequently, so will the self-report. For instance, it is possible that in station 1 students either did not have an appropriate illness script for back pain (or even pain in general) with which to approach this particular patient encounter or that the patient's history and physical examination in station 1 did not align well with students' illness script. Certainly, low levels of performance of critical action items in station 1 did indicate that students could not appropriately interview and examine the patient. Student difficulty performing (or knowing how to perform) appropriately in the station may have also precluded student ability to organize their thoughts into a detailed, accurate account of the clinical encounter in the patient note.

Differences in level of agreement between medical student and expert may also be connected to the sheer number of critical action items found in each station's behavioral checklist. There may have been a threshold of information that students could correctly recall and report in any given encounter. Station 2, the station in which students performed the best, had only 4 items and, therefore, may be under that threshold, while station 1 and station 3, with 7 items apiece, may be over. The complexity of a medical interview may lead to cognitive overload, in that the sheer amount of simultaneously-processed information overwhelms

a medical student's working memory. Past research has shown that SP ability to correctly report student behavior declines for longer checklists (Vu et al., 1992); it is also possible that students too had greater difficulty reporting a greater number of critical action items. Further study is required to determine the relationship between the quantity of critical items in a given station and student ability to correctly report performance of those items.

*Under-report in medical student self-report.* Medical students demonstrated a tendency to under-report their performance when compared to the expert rater, particularly physical examination items, resulting in scores that made them appear less competent; medical student training may explain this finding. While students may be adept at performing physical examinations, they do not necessarily know how best to report the physical examination in their patient notes. The physical examination in each station involved many parts, but the critical action items approach awarded credit for reporting only specific components of the physical examination. As novice physicians, training has focused on performing complete physical examinations and less on what specific maneuvers from a physical examination are most important to support a specific diagnosis. The problem for students was not that they did not perform certain maneuvers in a physical examination; rather, they did not know which components of a physical examination were important to reference given a particular diagnosis.

*Over-report in medical student self-report.* Students also reported on performance items that they had not, in fact, performed in the encounter, specifically information sharing items. This may also be due to medical student training. Though students were instructed to share with the patient a diagnosis and then report that shared diagnosis in their patient notes, many students offered the SP little concrete information about potential diagnoses in the

actual encounter, instead reporting the diagnosis only in the patient note. The information sharing aspect of a true clinical encounter, more often than not, and particularly in sensitive situations like those involving life-threatening illness, would likely be left to the physician. Medical students lack the experience and authority to perform such a task in an actual clinical encounter. A medical student member of the patient care team might interview a patient and then report findings to a physician in a patient note. Students, therefore, may have inexperience with sharing diagnoses with a patient, particularly life-threatening ones, and more experience with simply listing the diagnosis of that patient in the report of a clinical encounter. In a sense, over-report of information sharing may be encouraged, or reinforced, by the constraints of medical training in the clinics. Student level of performance of this behavior in the examination might be indicative of the kind of performance expected in an actual clinical setting and not what they were, in fact, instructed to do for the examination.

Alternatively there may have been a disconnect in perception of what students thought information sharing with the patient constituted. Medical students may be tailoring their communication to suit the listener, also known in the study of communication as recipient design, thereby speaking to the SP in a manner quite different from how they report the encounter to the faculty evaluator in the patient note. Medical students may have performed differently with the SP in the encounter, speaking in a manner they assumed was appropriate for a layperson, and reported in their patient notes in a completely different manner, using language they assumed appropriate for the faculty member who would later read and score the notes. Additionally, the desire by students to align themselves with the medical community of practice and gain social acceptance as legitimate members of that community may have prompted students to report information in a manner they felt would

be favorably viewed by an established practitioner (i.e., the faculty clinician), leading to social desirability bias in their self-reports. For instance, students may have thereby inadvertently over-reported their findings in their self-reports by not providing patients with specific diagnoses, incorrectly assuming such information was inappropriate or too difficult for a layperson to grasp in a short clinical encounter, and then provided the faculty member with a specific clinical diagnoses using language (i.e., medical jargon) they felt appropriate for a clinician. Students may have believed they had shared a clear description of a diagnosis with the patient (e.g., 'I'm concerned by your symptoms, and I would like to run some more tests to rule out the possibility of more serious conditions.'), but to the expert, this description was not specific enough. Students may not have appreciated how specific they needed to be with the SPs in the actual encounter to obtain credit for information sharing.

## 5.2 Improving medical student self-report

Improvement of medical student self-report depends on changes to the experience provided to students in the third-year of training, a time when students have the opportunity to gain real-world experience on a daily basis working with other physicians and members of the medical team and caring for actual patients. While the first two years of medical training (known as the pre-clinical years) build the medical interview skills of students and provide safe spaces for practice, using role play and SPs as well as physician mentors, the clinical years, and in particular, the third-year of training, may not provide ample opportunity for students to practice all necessarily skills equally, especially sharing information with the patient about diagnosis and treatment and the art of composing a detailed, specific, and

accurate reflection of care in the patient note. While much of this is likely due to the constraints of patient care, and the need to balance patient safety and privacy and the training of not just medical students but interns and residents as well, faculty should strive to make changes to the third year to provide students with increased experience and practice using these skills.

Faculty in each clinical rotation must provide good examples of patient notes, require students to complete patient notes, and also develop and implement a system that provides for meaningful feedback to students on their patient notes, specifically by others involved in the care of the same patients who are already familiar with the patients, and tracking of student progress. In other words, faculty need to develop and implement "public" ways of creating patient notes for medical students, both by modeling the composition of patient notes and by making patient notes composed by students more accessible for peer or faculty feedback by, for instance, creating student portfolios of notes based on each patient encounter, similar to teaching portfolios in education. Ensuring such an experience, however, is no easy feat, considering the constraints of physician time, limited opportunity for physician faculty development, and lack of medical student access to the patient electronic medical record (EMR) in the hospital system (Friedman, Sainte, & Fallar, 2010; Mintz et al., 2009). In many institutions, medical students cannot easily access the EMR, meaning they do not gain experience using an EMR nor can they easily access the designated space where physicians compose and place their own patient notes. Unless hospital policy is addressed, storing and tracking the progress of student patient notes may require the development of another, parallel system that collects information about patients, which can be difficult or even impossible to maintain, in light of patient security requirements, and confusing for

members of the patient care team. Faculty, however, must confront these issues in order to ensure that students receive training in correctly reporting patient information based on the clinical encounter.

In addition, faculty should consider investing in the development of self-report skills among practicing physicians who model these behaviors for students. Faculty must establish, firstly, how well interns, residents, and attending physicians themselves self-report clinical encounters and secondly, how often they model correct self-report for students. Improving faculty ability to correctly self-report and then providing them with the capability to demonstrate correct report to students may improve student self-report as well.

Finally, faculty may want to teach students memory recall skills when instructing students on how to complete a patient note. Given their level of training and experience with the patient note, students should perhaps, early in their training, be encouraged to report all information in their reports, no matter how mundane or unimportant it may appear, progressing over time to reporting only pertinent information in the note. A more experienced physician should be able to distill from a clinical encounter what information is important to report in the patient note. Perhaps students need to progress to this level gradually, first learning just to simply report, then learning how to pick and choose what information to report in the patient note. Reporting all information from a clinical encounter, however, is a daunting task. Correctly reporting back fifteen minutes of an encounter is no easy feat. Students may benefit from learning strategies to improve memory and recall, for instance, by learning how to unobtrusively take good notes in short hand during the actual patient encounter. In this way, students can learn how to report, and then learn how to select what to report, creating a pedagogical loop where by learning how to recall specific details of a

medical interview will allow students to further develop and hone their diagnostic skills.

*Improving self-report of physical examination.* Students participated in the CPX at the beginning of their fourth year of medical school, following a year of clinical rotations through various specialties. Clearly students were capable of performing the various maneuvers found in a complete physical examination, but they lacked the ability to correctly report all relevant details from those examinations. Students may require further training and, perhaps more importantly, experience based on actual patient encounters in the third year of training, not in how to perform a physical examination, which is introduced in the first year of training, but in how to determine what information from that examination is critical to report in the patient note.

One strategy might involve designating a physician or physicians on each patient care team to work directly with students on the correct report of physician examination in patient notes. Designated physicians would model with each student a physical examination and correct report of that physical examination and then observe each student conduct a physical examination on a patient with a known illness and review the student's patient note for that encounter. These faculty-physicians would need good self-report skills themselves, but also knowledge of common errors, like those reported in this study, found in medical student patient notes. Ultimately, such a strategy may require too much of physicians.

Another, perhaps less faculty time-intensive (and therefore more feasible) strategy, involves providing students with a module designed to develop student physical examination reporting skills, or student clinical reasoning specifically in relation to the physical examination. This could include written exercises such as script concordance tests, which assesses how well knowledge is organized when making clinical decisions (Brailovsky, Charlin, Beau-

soleil, Coté, & van der Vleuten, 2001; Charlin, Roy, & van der Vleuten, 2000; Fournier, Demeester, & Charlin, 2008), video review of a physician patient interview and review of a physician note of that encounter, as well as actual practice composing a patient note based on a video recording of a physician-patient encounter later evaluated by a faculty member, with specific attention paid to the correct report of the physical examination. Currently, the medical curriculum provides opportunity for the practice of physical examination maneuvers and for the composition of patient notes. Faculty in the clinical third year, when students build experience with actual patients, need to connect these two concepts to ensure that students are also practicing reporting what they actually performed.

*Improving self-report of information sharing.* In medical training, there may be a disconnect between student training on information sharing and student experience of information sharing in the actual clinics. During the first two years of medical training, students practice information sharing as part of the curriculum, using both role play with one another and SPs. However, in the third year, the constraints of actual patient care may preclude students from practicing this skill with actual patients. Students may rarely see an undifferentiated patient, or a patient whose illness is heretofore unknown, and talking to a patient about his or her illness may be left to the attending physician. Faculty must determine how often students perform this aspect of the clinical encounter with actual patients (and, in contrast, how often this is left to the attending physician). It may be that faculty who designed the CPX are attempting to capture a skill that fourth-year students have had little experience actually performing. Educators need to determine whether it is reasonable to require medical students to share information about diagnosis, a student behavior that is not reinforced in actual practice.

Information sharing, though, remains an important element of clinical care, and faculty should ensure that students are provided with some opportunity to practice this skill in their actual clinical experiences to ensure that what students report in their patient notes reflects what was discussed with patients during the encounter. Though students received training in how to share information with a patient, they may have lacked experience, especially during the third year, that reinforced and solidified that training. Faculty may consider implementing coaching sessions with students to encourage the sharing of information. For instance, students may interview a patient, meet briefly to discuss with their coach, and then, with the coach, talk with the patient about next steps in treatment. Students could then progress, with increased experience, to discuss elements of a patient's diagnosis with the patient under the tutelage of the coach. This process of legitimate peripheral participation, a theory introduced by Lave and Wenger (1991), allows students to gradually develop their information sharing skills, progressing from inexperienced medical students to experienced physician by gradually taking on more difficult tasks under the guidance of expert faculty-clinicians. While it might never be appropriate for medical students, who have relatively little training, experience, and authority in comparison to other members of the medical team, to reveal to a patient a diagnosis like HIV/AIDS, cancer or even, diabetes, a coaching relationship might provide students with the building blocks for future success sharing this information with patients and align student ability to report a diagnosis in the patient note with student ability to share that diagnosis in the actual encounter.

*Other solutions.* In addition to changes to medical student training, faculty should also consider changes to scoring that account for student inability to correctly self-report performance when using student self-report to evaluate competence. Possible strategies to explore

in further research include: a) adjustments to scores, such as weighting specific, more correctly reported critical action items (e.g., history) to account for under-reporting; or b) adjustments to how scores are constructed, for instance, by omitting from score construction certain kinds of data known not to be reported correctly (e.g., information sharing items) or by using alternate means to collect this data for the purposes of score construction; or c) adjustments to the criterion-referenced standard itself, for instance, by lowering the standard by which decisions are made regarding competence or by creating a new criterion-referenced standard altogether. While educators should consider primarily how to improve medical student training in completing of a patient note, faculty may also want to adjust scoring procedures to ensure fair and meaningful scores.

## 5.3 Areas of future research

*Other uses of self-report.* Though self-report may not be well positioned to score student clinical competence, it may prove useful in other areas, such as the study of medical student clinical reasoning, or student ability to arrive at a list of potential diagnoses. Self-report is a cornerstone of medical practice, used to communicate information pertinent to patient care between members of the medical team. Though not necessarily useful to the scoring of clinical competence, due to high levels of incorrect report, medical student self-report may play a role in other areas of medical education and assessment. Clinical reasoning, or the ability to sift through various pieces of information about patient history and physical examination findings to arrive at a list of likely diagnoses, is arguably one of the more important components of medical student clinical competence. It is important to note that

SPs cannot score medical student clinical reasoning as this is not (generally) an observable skill. Educators should consider examining medical student self-report for evidence of clinical reasoning, by, for instance, looking at the relationship between what information students provided about the patient and their diagnoses. What information do students provide when justifying certain diagnoses? How does this compare to a more experienced clinician's self-report? If self-report does provide an accurate reflection of student clinical reasoning, then faculty could use it for assessing and scoring this component of clinical competence, however this requires further study and exploration.

*Improving SP ratings.* Another important area of study, secondary to the current study, is examining why SPs had such difficulty correctly reporting student performance, considering their ratings form the basis for scoring medical student competence on the CPX. Possible research includes, for instance, stimulated recall exercises with SPs following the examination period to perhaps learn why certain items were so challenging to report correctly. It may also be worth investigating the success of different SP training interventions, such as additional practice for SPs in scoring items using video recordings of past examinations, and finally, changes to the behavioral checklist itself that improve the clarity of items. Such research could assist in the improvement of training of SPs and the use of their ratings to score student clinical competence.

## 5.4 Chapter summary

Student inability to correctly report performance may be linked to the third-year curriculum, as students may not have ample opportunity to practice certain elements of the clinical

encounter, like information sharing, and even less opportunity to compose and receive feedback on patient notes of these actual encounters. Though students gain much exposure to the skills needed to demonstrate clinical competence during their first and second years of training, as well as practice using these skills in simulated environments and during brief clinical encounters, they may not have ample exposure during the third, clinical year to reinforce that training. In addition, while the curriculum covers interviewing patients and reporting in a patient note, it is likely that the two are never analyzed and discussed in tandem. Rather than having these elements taught as two distinct skills, it may behoove faculty to connect the two, to ensure that students learn proper reporting skills.

Attention should be paid to bolstering not only student experience composing patient notes based on clinical encounters, but also timely faculty feedback on those notes with special attention paid to correct report. While faculty observation of student patient interviews and evaluation of patient notes of those encounter would easily address this deficiency in the curriculum, constraints of faculty time and the healthcare system make such practice a challenge. Alternatively, faculty can invest in alternate strategies to build student ability to make connections between what occurred in the patient encounter and what they report in their patient notes, using such tools as script concordance tests and simulated exercises using video review.

Future study should examine the use of self-report in other areas, not for scoring of clinical competence, but for, for instance, evaluating student clinical reasoning. Self-report is a key element of patient care and therefore requires attention in medical school curriculum. It also may present interesting future possibilities in the assessment and evaluation of medical student competence. While this study provided insight not only into the use of medical

student self-report for scoring of the CPX, it also is intended, more generally, to guide those considering the use of professional self-report in other contexts such as teacher evaluation and instructional improvement.

# Chapter 6

# Implications for teachers' use of self-report

## 6.1 Use of self-report in education

Teacher self-report takes many forms, including instructional logs, daily logs, and time diaries as well as teacher questionnaires and surveys, and is used for a variety of purposes, ranging from large-scale assessment of teacher instructional improvement programs (e.g., the Study of Instructional Improvement), to international study of teaching and learning (e.g., the Trends in International Mathematics and Science Study), smaller-scale, regional tools used to gain insight on teacher practices and attitudes for assessment purposes (e.g., the Texas Professional Development and Appraisal System), and even clinical evaluation of attention deficit hyperactivity disorder in students (e.g., the National Initiative for Children's Healthcare Quality Vanderbilt Assessment Scale). Instructional logs, daily logs, and time diaries require teachers to record teaching behaviors for a day (or shorter period) generally

using a highly or semi-structured form, organized either by kind of instruction (e.g., mathematics, literacy, etc.) and instructional practice (e.g., writing, reading comprehension, etc.) or by time of day. Researchers administer these instruments multiple times, over a specific period to capture a pattern of teacher practice. For instance, the Study of Instructional Improvement used teacher logs, as one of several tools, to collect information about daily teacher practices to assess the impact of reform programs (Rowan, Camburn, & Correnti, 2004). The American Time Use Survey, which is administered by the U.S. Bureau of Labor Statistics to a sample of Americans, including teachers, uses time diaries to collect information about teacher practices to establish teacher work patterns over the course of a year and compare those work patterns to those of other professions (U.S. Bureau of Labor Statistics, 2013). Teacher surveys and questionnaires can also be used to measure specific constructs, such as evaluation of the implementation of a new curriculum, assessment of a particular student or students, or teacher reflection on policies and school environment. These may require teachers to retroactively report on performance over the course of a much longer period of time, such as the entire academic year. The National Assessment of Educational Progress, which assesses knowledge and performance of American children in certain subject areas, uses teacher questionnaires to give more context to student performance by collecting information on teacher training and instructional practices (National Center for Education Statistics, 2012).

Teacher self-report is currently being used to make important decisions such as promotion and tenure, the allocation of funding, as well as being used in studies of instructional improvement, which attempt to link reported behaviors by teachers to student outcomes. Teacher self report can be used to assess teacher fidelity to a curriculum, establish best prac-

tices, and offer opportunities for teacher reflection and change. Self-report is well-suited to these endeavors for several reasons. Firstly, teacher self-report, particularly questionnaires and surveys, is often far less costly to implement than independent observation (Camburn & Barnes, 2004; Rowan et al., 2004). Secondly, use of a variety of methods allows for triangulation of teaching measures and a more likely accurate representation of teacher performance (Dwyer, 1994). Thirdly, some argue that the teacher is in the best position to accurately document what he or she actually did in the context of his or her own classroom (Camburn & Barnes, 2004) and can best capture the level of preparation and expertise he or she brought to the class (Kulik & McKeachie, 1975). Finally, some forms of self-report like logs and time diaries can better capture the complexity of teaching than can observation or video review, which oftentimes rely on a small sample of teaching events in order to generalize teacher practice even though teaching activities, content, and instruction can vary considerably from day to day, week to week, and month to month (Rowan & Correnti, 2009). Despite this interest in the use of self-report, self-report methods do not receive much attention in the education literature (Rowan et al., 2004).

Self-report, however, is not without its limitations, mainly stemming from teacher misperception of teaching effectiveness (Blackburn & Clark, 1975; Centra, 1982) and potential inaccuracy in teacher memory, particularly when using year-end surveys or questionnaires (Rowan & Correnti, 2009). Previous research is inconsistent about whether self-report data among elementary and high school teachers agree with other sources of information, such as expert raters' observations (Ball et al., 1999; Hardebeck et al., 1974; Koziol & Burns, 1986; Newfield, 1980; Rowan & Correnti, 2009; Camburn & Barnes, 2004). On the one hand, Rowan and Correnti (2009) found good levels of agreement between teacher self-report in

logs and third party observation of teacher literacy instruction, averaging 75% on specific details of behaviors tied to the instructional focus of teaching reported in the log. On the other hand, Wheeler and Knoop (1982) found significantly different ratings between student teachers' self-reports and supervisors' ratings of performance. It is possible that disagreement between a teacher and an expert observer in the report of performance could stem from both under- and over-reporting. Below, I use two case studies–the Study of Instructional Improvement Language Arts teacher log and the Texas Professional Development and Appraisal System teacher self-report–to discuss possible sources of teacher under-reporting and over-reporting, as well as strategies for combating this tendency.

## 6.2 Accuracy of teacher self-report

*Case study: Study of Instructional Improvement (SII) Language Arts Teacher Log.* The Study of Instructional Improvement (SII) investigated the impact of three comprehensive school reform programs (Accelerated Schools Project, America's Choice, and Success for All) on teacher instruction and student achievement in 112 elementary schools with high levels of poverty (Rowan et al., 2004; Rowan & Correnti, 2009). Over a four-year period, University of Michigan researchers used a variety of tools to collect data from approximately 2000 teachers as well as school administrators, students, and parents. One such tool was the Language Arts teacher log, which consisted of over 100 dichotomously scored items related to teacher instruction during one day with one student; the log is found on the project's website (www.sii.soe.umich.edu/instruments/). Teachers who participated in the study received training and ongoing support in completing three extended periods of logging

over the course of one year. During each logging period, teachers reported on instruction that day with a sample of 8 students, or "target" students, rotating their reports by student. Broad topics of literacy instruction focus included: comprehension, writing, word analysis, concepts of print, reading fluency, vocabulary, grammar, spelling, and research strategies.

The Language Arts Teacher Log asked teachers to report on their performance while interacting with the target student on that day, specifically focusing in greater detail, when relevant, on work in three sub-areas: comprehension, writing, and/or word analysis. Within each of these sub-areas, teachers reported whether or not students had performed certain specific tasks, whether the teachers themselves had demonstrated specific behaviors, and how teachers responded to or evaluated student performance. These lists of performance items were highly detailed and extensive.

It is important to note that the SII language arts teacher self-report captured frequency, over time, of certain behaviors in the classroom, not quality of instruction. Later, project researchers aimed to identify the relationship between these behaviors and student achievement outcomes. In the current study, using the critical action items approach to scoring, faculty made decisions about medical student competence based on the quantity of items performed by the students. Faculty interpretation of the frequency of items performed, and not medical student self-report itself, determined clinical competence, meaning self-report as used in the current study also did not purport to capture quality of performance. Capturing quality using self-report remains an important issue to teacher self-report (Hamilton & Martinez, 2007), though it is not directly addressed by this study.

Findings of the current study suggest two areas where teachers may be prone to incorrect report of performance on the SII teacher log: 1) in report of what teachers demonstrate to

their students (by under-reporting); and 2) in report of how the teachers respond to and interpret student performance (by over-reporting). Like medical student report of physical examination maneuvers, which require students to perform an action on an SP and then report SP response, teachers may have under-reported their performance of specific behaviors. For instance, teachers could have under-reported demonstrating a specific reading comprehension skill or think-aloud writing exercise. Medical students under-reported specific elements of their physical examinations, despite completing entire examinations on the patient. Likewise, teachers may have omitted from their reports particular actions from a much larger lesson or interaction with the student that involved multiple behaviors to teach comprehension and writing. Teachers may omit their own practices not necessarily because they forget what activities they have completed with students, but because they do not consider these activities, or even the minutiae of these activities, important to report despite their importance to those conducting the research. Interpretation by teachers of student performance could be subject to over-reporting in the SII project. Medical students demonstrated a tendency to over-report information sharing items, meaning they failed to share with the patient information about his or her condition in the encounter though they reported it in the patient note. Likewise, teachers in the SII project might have over-reported what they shared with students about their performance, like, for instance, telling students how they might improve their writing or correcting students who made errors when performing word analysis. Again, teachers may not be attempting to misrepresent themselves; rather, they could report with greater clarity or specificity their feedback to students compared to their actual performance in the teacher-student encounter.

Additionally, teachers may report more correctly in some content domain areas than

others. The SII teacher log asked in greater detail about three focus areas of instruction: comprehension, writing, and word analysis. It is possible that teachers reported with greater accuracy, for instance, word analysis and writing, but not comprehension. Of all three domain areas, comprehension contained the most items that depended on or specifically asked for interpretation of student performance, which, as described above, may be particularly susceptible to over-reporting.

Interestingly, study of the teacher log by SII project researchers supported the fact that disagreement can arise between teachers and observers in the report of performance, and that certain features of the context and content of teaching may partially explain discrepancies. Camburn and Barnes (2004) conducted a study comparing data provided by 31 teachers (of the nearly 2000 in the study) in their logs (using an earlier iteration of the log) of one day of instruction to two SII trained-observer logs of that same day. It should be noted that the study did not treat disagreement as instances of "incorrect" report by the teacher; rather, the authors sought to understand and explain how disagreement arose, casting blame neither on the teacher nor the observer. Though the researchers did use a slightly different version of the log, which consisted of nearly 50 more items than later versions, there were some general conclusions made by the authors that corroborate the potential areas of incorrect report suggested by the present study. The authors found that teachers and observers disagreed more about comprehension and writing than word analysis items. They also disagreed less about reporting student performance of tasks, or "student activity." Also striking, the authors found that teachers and observers agreed proportionately more on items that happened more frequently, in other words, less "difficult" items. The finding is consistent with the present study's finding that medical student incorrect report was greater for more

difficult stations.

*Case study: the Texas Region 13 Professional Development and Appraisal System (PDAS) Teacher Self-Report.* Created by the Texas state legislature, the Education Service Center Region 13 is one of many public centers that strives to provide educational services to school districts in its region by providing support–such as professional development, training, and instructional resources–for teachers and school administrators. Information about Region 13 is available on their website (http://www4.esc13.net/). The Professional Development and Appraisal System (PDAS) was, up until until 2012, the Texas state approved instrument for evaluating teacher performance and identifying areas in need of professional development. The PDAS process included one 45-minute observation and a teacher self-report form, which consisted of 9 short response questions regarding teacher instructional practices. The form can be found on the Region 13 website (http://www4.esc13.net/pdas/). New teachers to the region were required to complete training on the PDAS before completing the self-report.

Though the PDAS self-report questionnaire used some strategies to encourage correct report, some prompts, based on the results of the present study, may have yielded incorrect report. The format of the PDAS was much less structured than the SII teacher log, providing teachers with short answer prompts, thereby forcing teachers to rely heavily on correct recall of relevant details of instruction. Self-report prompts included: academic skills taught by teachers, instructional adjustments made by teachers, approaches to monitoring classroom progress and providing students with feedback, strategies for dealing with truant and failing students, professional development activities, the impact of those professional development activities on the classroom, and areas in need of improvement.

There are multiple places where teachers may have over-reported their activities, including ways in which they had made adjustments to their teaching based on assessment of student achievement and how they provided feedback to their students regarding performance. Like information sharing, description of adjustments to teaching and feedback provided to students requires teachers to interpret student performance and "diagnose" the problem in order to make meaningful changes. Teachers may over-report adjustments they have made or ways in which they provided feedback to students about their ongoing performance. Perhaps teachers may report the specific feedback they intended to provide students in their self-report, but observation might reveal this feedback was actually much more vague or non-specific. In particular, the PDAS asked teachers to describe all approaches they used to provide feedback to students. Requesting all approaches may have lead some teachers to over-report ways in which they provided feedback to students in an effort to appear especially competent. In other words, the format of the PDAS self-report may lead to social desirability bias, or the tendency for teachers to report what they think evaluators and administrators want to see–in the case of the PDAS self-report, to over-report instructional activities–and not based on what they truly believe about their performance. Likewise, medical students may have communicated with the patient in a manner they felt befitted a layperson but then reported what they thought a faculty evaluator or grader would want to see in a patient note, not realizing the two different versions of their performance did not completely align. Social desirability bias is a serious issue for researchers using self-report and can easily interfere with the interpretation of results. For education researchers, as well, social desirability bias remains an important concern for any researcher designing a teacher self-report instrument, particularly if results may be tied to teacher evaluation or pay.

There are also possible sources of under-reporting. Teachers using the PDAS may have also exhibited a tendency to under-report behaviors used to motivate struggling students (e.g, often truant, failing, etc.), much like medical students oftentimes under-reported elements of the physical examination. Two of the nine questions included on the self-report questionnaire asked teachers to describe actions they had taken or their approach to students who were experiencing difficulty in the classroom. These prompts did not limit responses to one example, meaning teachers could easily list some approaches but not all approaches, or, even, omit some actions that would have been included as student assistance by another teacher or observer.

Although use of the PDAS has ended within Texas due to pressure to develop improved teacher assessment systems, the implications discussed here may apply to other assessment systems. Based on the findings of this study, it is very possible that the PDAS and other similar self-report questionnaires could be improved to facilitate the collection of correct information.

## 6.3 Strategies for improving teacher self-report

There are a variety of strategies that might be considered for improving the accuracy of teacher self-report. General strategies to consider include reliance on more structured self-report formats (e.g., survey) versus less structured formats (e.g., free response), clear instructions and training in the use of the self-report tool, and tool piloting, playing close attention how well teachers capture their performance.

*Case study: Study of Instructional Improvement Language Arts teacher log.* Though SII project researchers strove to provide teachers with an exhaustive list of possible teaching behaviors, these very descriptions of teacher instruction may have confused teachers completing the log. The SII project teacher log consisted of around 100 detailed individual items that broke teacher instruction down into distinct, finite behavioral items. Researchers may have used language in the log that was more familiar or accessible to the researchers than to the actual teachers. In fact, follow-up interviews conducted by Cambrun and Barnes (2004) revealed that teachers and observers interpreted certain terms used in an earlier version of the log differently, despite efforts to explain to teachers the meaning of terms during training and in the teacher log use manuals. One strategy when creating such detailed logs may be to use teachers' language or description of behaviors, gleaned during informational interviews with teachers about instructional practices, to create the wording of individual log items. Aligning the language of the teacher log with the language of teachers may improve teacher ability to correctly complete the log.

Changes to the length and organization of the log may also improve teacher self-report. At approximately 100 items, the log is an exhaustive list of teacher instructional behaviors. Medical students reported a significantly higher proportion of correct critical action items in the one station with the least number of critical action items. SPs also have been shown to report more accurately using shorter checklists (Vu et al., 1992). The sheer length of the log may have impeded teachers from correctly reporting behaviors. The SII project researchers did shorten the length of the log with the final version of the log becoming shorter by nearly 50 items, though the log remains long. In addition, organization of these 100 items could potentially assist teachers in correctly self-reporting their practices. Like use

of teacher language, how do teachers describe the structure of their instruction? Organizing the teacher log around teachers' own organization of the learning encounter, could help teachers more effectively recall behaviors. SII project researchers did make subtle changes to the organization of the log over time. The final version of the log still consisted of the three main categories–comprehension, writing, and word analysis–but also had sub-groupings of items within each of these categories. Ascertaining whether or not this organization aligns well with teacher perception of organization of instruction could further improve teacher ability to report instruction.

*Case study: the Texas Region 13 Professional Development and Appraisal System (PDAS) teacher self-report.* Changes to the PDAS self-report (or similar instruments) could improve teacher ability to report on their performance. The PDAS instrument asked teachers to retrospectively recall and report cumulative behaviors. The open-ended questions may have proven too difficult for teachers to respond to with much accuracy. Firstly, the format assumes that all involved–the evaluators, supervisors, and teachers–have similar concepts of what constitutes an "approach" to providing feedback to students or addressing struggling students. A teacher may list an approach that is not considered valid by evaluators or may conversely omit a practice that would have been considered valid thinking it inconsequential or irrelevant. Secondly, open-ended questions fall short of stimulating teacher recall, placing all responsibility to recall all possible relevant information squarely on the teacher. One strategy to improve teacher self-report, then, is to ask more focused questions, such as asking for report of one example of a particular approach or dealings with one struggling student, or asking teachers for an example of one approach that worked well (and one that did not). Thirdly, the PDAS self-report, and instruments like it, asked teachers to retrospectively re-

port cumulatively on activities that had occurred over a considerable length of time, unlike teachers participating in the SII project or medical students composing patient notes in the current study, both of whom had the advantage of reporting on activities that had occurred over a relatively shorter period of time (one day or 15 minutes prior, respectively). Past research suggests that self-report becomes less accurate when respondents are asked to reflect over a longer period of time, like a school year (Rowan et al., 2004). One strategy to combat this is confining teacher response to actions that occurred during a short time span (e.g., that day or the day preceding). The simple fact that results from the current study demonstrated that even when reporting on behaviors performed 15 minutes prior, medical students were still incorrect in their reports only underscores the importance of timely self-report.

Lastly, ongoing training in use of the self-report tool could also improve teacher self-report. Teachers new to the region were required to complete training in the PDAS; however, there is no evidence of ongoing support and training. Like medical students, teachers using self-report may benefit not just from training in instructional methods but in how to correctly report performance. This training should be reinforced with opportunities to practice and to receive formative feedback on their self-report in order for teachers to gain experience providing correct self-report.

## 6.4 Areas for future study

Teacher self-report can serve an important role in teacher evaluation and professional development; however, education researchers desiring to implement a self-report as part of a study must determine what contexts and conditions best facilitate correct self-report by

teachers. This includes investigation of teacher self-report of different instructional practices in different learning contexts. For instance, a future study should examine how teacher self-report of specific instructional practices in certain learning experiences compares to report by an independent, third-party expert observer. Such a study might examine teacher self-report by piloting the self-report instrument immediately following a reading comprehension, writing, and word analysis lesson in three domains: student activity, teacher instructional practices, and evaluation and feedback to students. An expert observer would later watch a video recording of the same encounter, noting the presence of absence of certain behaviors of interest in these three domains. Later, researchers could analyze the match between teacher and observer, identifying instances of under- and over-report by the teacher, and determining whether the kind of information reported or the context of the lesson explained variation in agreement. Results could then be used to refine training in use of the self-report or change the language and/or structure of the self-report instrument itself, resulting (hopefully) in improved teacher self-report in the study.

Another important area of investigation for education researchers is self-report design. Within medicine, physician self-report generally follows a standard format and is completed at regular, predictable intervals (though small variation can exist from institution to institution). However, within education research, as indicated in this chapter by the use of two case studies each involving a very different kind of self-report, there is a wide variety of possible teacher self-report formats administered at very different times during the routine practice of the teachers involved. Which format–teacher log, time diary, survey, questionnaire, etc.–and over what span of time span of time–by lesson, daily, weekly, by semester, annually–best promotes correct self-report among teachers? Investigation by Rowan and Correnti (2010)

demonstrated that teachers tended to over-report the frequency of behaviors in an annual survey in comparison to the summed totals from daily teacher logs. Researchers desiring to use self-report in a study should consider implementing more than one strategy of self-report as well as at least one instance of third-party observation to determine which method of self-report yields the most accurate information about teacher practices. Additionally, attention must be paid to the construction, language, and organization of the self-report, using teacher focus groups, instrument piloting with a small group of teachers, and follow-up with teachers who have used the self-report. These strategies may help improve the overall use and accuracy of data gleaned from teacher self-report.

The present study highlights potential sources of inaccuracy in information when using teacher self-report and suggests some strategies for improving level of information accuracy, with the ultimate aim of assisting researchers attempting to incorporate this tool into their own research studies. Teacher self-report has unique advantages. Self-report collects information about instructional practice from an active participant in the classroom, while oftentimes costing less than third party observation. Logistically, teacher self-report may also be easier to implement and less disruptive to teaching than peer, supervisor, or third-party observations. Proper use of teacher self-report, however, requires researchers to consider potential sources of incorrect report among teachers in the study as well as to develop and implement strategies to ensure the quality of data. Education researchers can benefit from this study of the deep-rooted practice of self-report in the medical profession to provide insight on the use of self-report in teacher education, training, evaluation, and development.

# Chapter 7

# Conclusion

Use of self-report to collect data on professional practice is increasingly widespread with self-report playing an important role in professional assessment and development. Educators and researchers across professions find self-report–in the form of logs, time diaries, questionnaires, and surveys–useful for a variety of purposes, including: determining competence, assessing the implementation of techniques, and identifying and reflecting on particular professional behaviors. Logistically, implementing a self-report can cost less, reach a wider audience, and prove less taxing and disruptive to standard professional practice and the environment in which it takes place, all features that make self-report a desirable methodological tool.

Self-report appears in its various forms across professions, including medicine and education. In the field of medicine, self-report plays an important role in medical training. The ability to compose a succinct, accurate, and informative patient note is a lifelong skill for any physician. Among teachers, it is not evident that teachers necessarily receive instruction in the use of self-report as part of training. However, self-report often is used

by researchers and administrators to assess teacher instructional practices, monitor implementation of curriculum, evaluate student performance (or, rather, teacher interpretation of student performance), and inform teacher development.

Given the widespread use of self-report, researchers must determine how well professionals can effectively use self-report to produce accurate information about professional practice, for instance, in comparison to other methods like third-party observers, peers, and supervisors. In this vein, this study addressed whether or not self-report could serve as a valid substitute for an expert observer's judgment about competence in carrying out professional tasks.

This study revealed that professional self-report did not agree highly with an expert observer. The professional did not always accurately report what he or she had done or not done, sometimes over-reporting behaviors, sometimes under-reporting behaviors. Though past research has suggested such findings could indicate potential falsification (i.e., cheating), other explanations like incomplete training in the use of self-report, instrument wording, organization, and length, and even the intricacies of human cognition could also equally account for the discrepancies between a professional's self-report and an expert observer's documentation of an encounter.

Given these findings, one readily apparent implication is that it is certainly risky to assume that self-report can be an accurate reflection of practice, and strategies must be employed to improve accuracy. Performance of behaviors and report of those behaviors appear to be two distinct and separate tasks, each of which requires some development in the professional. This study did reveal that in certain contexts and for certain kinds of information, professionals were able to self-report correctly their behaviors. Attention must be paid to the ongoing improvement of self-report accuracy across desired contexts

and for desired practices in order to ensure the validity of data and the interpretation of self-report data. Accuracy of self-report may be improved by treating self-report not as an innate skill but rather as a skill that requires training, practice, and ongoing development in the professional. Such training can take many forms including coaching and mentoring as well as workshops and training modules and can even include training in memory and recall. Training should also include provisions for ongoing assessment and evaluation of professional use of self-report to ensure that standards of correct report are maintained. What is more, developers of self-report instruments must determine how their tools impact (potentially negatively) opportunities for correct self-report by, for instance, piloting the instrument with intended users, conducting focus groups, and performing qualitative research to determine how professionals make use of the self-report and whether or not that use aligns with what was intended. Understanding the professional's thought-processes and sense-making can aid in the development of strategies for improving ability to self-report.

Improving self-report accuracy not only benefits researchers and administrators who make use of the tool; it also has the potential to positively benefit professionals. Self-report allows researchers to gather information about a learning encounter from an actual participant in that encounter, be it a physician-patient clinical encounter or a teacher-student encounter, often for a fraction of the cost of third-party raters and without infringing on the time and energies of supervisors, peers, and administrators. In addition to collecting information about teacher practices, self-report also allows for the collection of data from teachers about how they interpret and process student activity and performance. Improved self-report skills may lead to greater awareness of and improved insight into the professional's own practice, which, in turn, may be an important condition for the ongoing improvement of practice.

Only correct self-report can allow for the identification of skills in need of improvement as well as skills in which the professional already excels. For instance, medical students may not need increased training in counseling a patient on smoking cessation. Perhaps they simply need more training on how to document this practice in the patient note. Likewise, teachers may not be aware that certain practices positively (or negatively) influence student achievement. Correct self-report allows teachers to make connections between their practice and their students, potentially benefiting both teacher development and student outcomes. Though its accuracy may be initially mistrusted, professional self-report is too ubiquitous and its potential advantages too attractive to ignore. It is therefore of utmost importance to ensure the ability of self-report tools to accurately capture true performance.

# Bibliography

Ball, D. L., Camburn, E., Correnti, R., Phelps, G., & Wallace, R. (1999). *New tools for research on instruction and instructional policy: a web-based teacher log* (A CTP Working Paper No. W-99-2). Seattle, WA: University of Washington, Center for the Study of Teaching and Policy. Retrieved from `https://depts.washington.edu/ctpmail/PDFs/Teacher_Log.pdf`

Barrows, H. S., Williams, R. G., & Moy, R. H. (1987). A comprehensive performance-based assessment of fourth-year students' clinical skills. *Journal of Medical Education*, *62*, 805–809.

Beaulieu, M., Rivard, M., Hudon, E., Saucier, D., Remondin, M., & Favreau, R. (2003). Using standardized patients to measure professional performance of physicians. *International Journal for Quality in Health Care*, *15*, 251–259. doi: 10.1093/intqhc/mzg037

Berwick, D. M. (2002). A user's manual for the IOM's 'Quality Chasm' report. *Health Affairs*, *21*, 80–90. doi: 10.1377/hlthaff.21.3.80

Blackburn, R. T., & Clark, M. J. (1975). An assessment of faculty performance: some

correlates between administrator, colleague, student and self-ratings. *Sociology of Education*, *48*, 242–256.

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

Brailovsky, C., Charlin, B., Beausoleil, S., Coté, S., & van der Vleuten, C. (2001). Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. *Medical Education*, *35*(5), 430–436. doi: 10.1046/j.1365-2923.2001.00911.x

Brannick, M. T., Erol-Korkmaz, H. T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores. *Medical Education*, *45*, 1181–1189.

Camburn, E., & Barnes, C. A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal*, *105*(1), 49–73.

Centra, J. A. (1982). *Determining faculty effectiveness: Assessing teaching, research, and service for personel decisions and improvement.* San Francisco, CA: Jossey-Bass Publishers.

Chambers, K. A., Boulet, J. R., & Furman, G. E. (2001). Are interpersonal skills ratings influenced by gender in a clinical skills assessment using standardized patients? *Advances in Health Sciences Education: Theory and Practice*, *6*, 231–241.

Charlin, B., Roy, L., & van der Vleuten, C. B. F. G. C. (2000). The Script Concordance test: a tool to assess the reflective clinician. *Teaching and Learning in Medicine: An International Journal*, *12*(4), 189–195.

Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research [Essay]. *Academic Medicine*, *75*, 182–190.

Cohen, R. J., Ek, K., & Pan, C. X. (2002). Complementary and alternative medicine (CAM) use by older adults: a comparison of self-report and physician chart documentation. *The Journals of Gerontology: Series A Medical Sciences*, *57*, 223–227. doi: 10.1093/gerona/57.4.M223

Colliver, J. A., Vu, N. V., Marcy, M. L., Travis, T. A., & Robbs, R. S. (1993). Effects of examinee gender, standardized-patient gender, and their interaction on standardized patients' ratings of examinees' interpersonal and communication skills. *Academic Medicine*, *68*, 153–157.

de Champlain, A. F., MacMillan, M. K., Margolis, M. J., King, A. M., & Klass, D. J. (1998). Do discrepancies in standardized patients' checklist recording affect case and examination mastery-level decisions? *Academic Medicine*, *73*(Suppl. 10), S75–S77.

de Champlain, A. F., Margolis, M. J., King, A., & Klass, D. J. (1997). Standardized patients' accuracy in recording examinees' behaviors using checklists. *Academic Medicine*, *72*(Suppl. 10), S85–S87.

Doyle, K. O., & Crichton, L. I. (1978). Student, peer, and self evaluations of college instructors. *Journal of Educational Psychology*, *70*, 815–826. doi: 10.1037/0022-0663.70.5.815

Doyle, K. O., & Webber, P. L. (1978). Self ratings of college instruction. *American Educational Research Journal*, *15*, 467–475. doi: 10.3102/00028312015003467

Dresselhaus, T., Luck, J., & Peabody, J. (2002). The ethical problem of false positives: a prospective evaluation of physician reporting in the medical record. *Journal of Medical Ethics*, *28*, 291-294. doi: 10.1136/jme.28.5.291

Dwyer, C. A. (1994). Criteria for performance-based teacher assessments: validity, standards, and issues. *Journal of Personnel Evaluation in Education*, *8*, 135–150.

Eastman, A. (1970). How visitation came to Carnegie-Mellon University. In K. E. Eble (Ed.), *The Recognition and Evaluation of Teaching* (pp. 75–89). Salt Lake City, UT: Project to Improve College Teaching.

Ellis, P. M., Blackshaw, G., Purdie, G. L., & Mellsop, G. W. (1991). Clinical information in psychiatric practice: what do doctors know, what do they think is known and what do they record? *Medical Education*, *25*, 438–443.

Fournier, J. P., Demeester, A., & Charlin, B. (2008). Script concordance tests: guidelines for construction [Correspondence]. *BMC Medical Informatics and Decision Making*, *8*, 18–24. doi: 10.1186/1472-6947-8-18

Friedman, E., Sainte, M., & Fallar, R. (2010). Taking note of the perceived value and impact of medical student chart documentation on education and patient care. *Academic Medicine*, *85*, 1440–1444. doi: 10.1097/ACM.0b013e3181eac1e0

Govaerts, M. J., van der Vleuten, C. P., Schuwirth, L. W., & Muijtjens, A. M. (2007). Broadening perspectives on clinical performance assessment: rethinking of the nature of in-training assessment. *Advances in Health Science Education: Theory and Practice*, *12*, 239–260. doi: 10.1007/s10459-006-9043-1

Hamdy, H., Prasad, K., Anderson, M. B., Scherpbier, A., Williams, R., Zwierstra, R., & Cuddihy, H. (2006). BEME systematic review: predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher*, *28*, 103–116. doi: 10.1080/01421590600622723

Hamilton, L. S., & Martinez, J. F. (2007). What can TIMSS surveys tell us about mathematics reforms in the United States during the 1990s? In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 127–174). Washington, D. C.: Brookings Institution Press.

Han, J. J., Kreiter, C. D., Park, H., & Ferguson, K. J. (2006). An experimental comparison of rater performance on an SP-based clinical skills exam. *Teaching and Learning in Medicine: An International Journal*, *18*, 304–309. doi: 10.1207/s15328015tlm1804_5

Hardebeck, R. J., Ashbaugh, C. R., & McIntyre, K. E. (1974). *Individualization of instruction by vocational and non-vocational teachers* (No. ED 131 202). Austin, TX: University

of Texas at Austin, Department of Educational Administration, The Office of School Surveys and Studies.

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.

Hartman, S. L., & Nelson, M. S. (1992). What we say and what we do: self-reported teaching behavior versus performances in written simulations among medical school faculty. *Academic Medicine*, *67*, 522–527.

Henry, B. W., & Smith, T. J. (2010). Evaluation of the FOCUS (Feedback On Counseling Using Simulation) instrument for assessment of client-centered nutrition counseling behaviors. *Journal of Nutrition Education and Behavior*, *42*, 57–62. doi: 10.1016/j.jneb.2008.12.005

Howard, G. S., Conway, C. G., & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology*, *77*(2), 187–196. doi: 10.1037/0022-0663.77.2.187

Kintsch, W., & Greene, E. (1978). The role of culture-specific schemata in the comprehension and recall of stories. *Discourse Processes*, *1*, 1–13. doi: 10.1080/01638537809544425

Kopp, K. C., & Johnson, J. A. (1995). Checklist agreement between standardized patients and faculty. *Journal of Dental Education*, *59*, 824–829.

Koziol, S. M., & Burns, P. (1986). Teachers' accuracy in self-reporting about instructional practices using a focused self-report inventory. *The Journal of Educational Research*, *79*, 205–209.

Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. *Review of Research in Education*, *3*, 210–240.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge, United Kingdom: Cambridge University Press.

Levenstein, J. H., McCracken, E. C., McWhinney, I. R., Stewart, M. A., & Brown, J. B. (1986). The patient-centered clinical method. 1. A model for the doctor-patient interaction in family medicine. *Family Practice*, *3*, 24–30.

Luck, J., & Peabody, J. W. (2002). Using standardised patients to measure physicians' practice: validation study using audio recordings. *BMJ*, *325*, 679–682. doi: 10.1136/bmj.325.7366.679

Luck, J., Peabody, J. W., Dresselhaus, T. R., Lee, M., & Glassman, P. (2000). How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. *The American Journal of Medicine*, *8*, 642–649.

MacMillan, M. K., Fletcher, E. A., de Champlain, A. F., & Klass, D. J. (2000). Assessing post-encounter note documentation by examinees in a field test of a nationally administered standardized patient test. *Academic Medicine*, *75*, S112–S114.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: a comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology*, *71*, 149–160. doi: 10.1037/0022-0663.71.2.149

May, W. (2008). Training standardized patients for a high-stakes clinical performance examination in the California Consortium for the Assessment of Clinical Competence. *The Kaohsiung Journal of Medical Sciences*, *24*, 640–645. doi: 10.1016/S1607-551X(09)70029-4

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

Miller, J. B., deWinstanley, P., & Carey, P. (1996). Memory for conversation. *Memory*, *4*, 615–631. doi: 10.1080/741940999

Mintz, M., Narvarte, H. J., O'Brien, K. E., Papp, K. K., Thomas, M., & Durning, S. J. (2009). Use of electronic medical records by physicians and students in academic internal medicine settings. *Academic Medicine*, *84*, 1698–1704. doi: 10.1097/ACM.0b013e3181bf9d45

National Center for Education Statistics. (2012, September 9). *NAEP–Overview.* Retrieved from `http://nces.ed.gov/nationsreportcard/about/`

Newble, D. I., & Swanson, D. B. (1988). Psychometric characteristics of the objective structured clinical examination. *Medical Education*, *22*, 325–334.

Newfield, J. (1980). Accuracy of teacher reports: reports and observations of specific classroom behaviors. *The Journal of Educational Research*, *74*, 78–82. Retrieved from http://www.jstor.org/stable/27507243

O'Hanlon, J., & Mortensen, L. (1980). Making teacher evaluation work. *The Journal of Higher Education*, *51*, 664-672. Retrieved from http://www.jstor.org/stable/1981171

Pangaro, L. N., Worth-Dickstein, H., MacMillan, M. K., Klass, D. J., & Shatzer, J. H. (1997). Performance of "standardized examinees" in a standardized-patient examination of clinical skills. *Academic Medicine*, *72*, 1008–1011.

Payne, N. J., Bradley, E. B., Heald, E. B., Maughan, K. L., Michaelsen, V. E., Wang, X. Q., & Jr., E. C. C. (2008). Sharpening the eye of the OSCE with critical action analysis. *Academic Medicine*, *83*, 900–905. doi: 10.1097/ACM.0b013e3181850990

Peabody, J. W., Luck, J., Glassman, P., Dresselhaus, T. R., & Lee, M. (2000). Comparison of vignettes, standardized patients, and chart abstraction: A prospective validation study of 3 methods for measuring quality. *Journal of the American Medical Association*, *283*, 1715–1722. doi: 10.1097/ACM.0b013e3181850990

Pieters, H. M., Touw-Otten, F. W. W. M., & de Melker, R. A. (2002). Simulated patients in assessing consultation skills of trainees in general practice vocational training: a validity study. *Medical Education*, *28*, 226–233. doi: 10.1111/j.1365-2923.1994.tb02703.x

Richter Lagha, R. A., Boscardin, C. K., May, W., & Fung, C. (2012). A comparison of

two standard-setting approaches in high-stakes clinical performance assessment using generalizability theory. *Academic Medicine*, *87*, 1077–1082.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, *37*, 322–336. doi: 10.1037/0022-3514.37.3.322

Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: a study of literacy teaching in third-grade classrooms. *The Elementary School Journal*, *105*, 75–101. doi: 10.1086/428803

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: lessons from the study of instructional improvement. *Education Researcher*, *38*, 120–131. doi: 10.3102/0013189X09332375

Stafford, L., & Daly, J. A. (1984). Conversational memory: the effects of recall mode and memory expectancies on remembrances of natural conversation. *Human Communication Research*, *10*, 379–402. doi: 10.1111/j.1468-2958.1984.tb00024.x

Stilson, F. R. B. (2009). *Psychometrics of OSCE standardized patient measurements.* (Doctoral dissertation). Retrieved from `ProQuest Digital Dissertations Database. (AAT 3420523)`

Szauter, K. M., Ainsworth, M. A., Holden, M. D., & Mercado, A. C. (2006). Do students do what they write and write what they do? The match between the patient encounter and patient note. *Academic Medicine*, *81*(Suppl. 10), S44–S47.

Tamblyn, R. M. (1989). *The use of the standardized patient in the measurement of clinical competence: The evaluation of selected measurement properties.* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations Database. (AAT NL57157)

Tamblyn, R. M., Abrahamowicz, M., Dauphinee, W. D., Hanley, J. A., Norcini, J., Girard, N., ... Brailovsky, C. (2002). Association between licensure examination scores and practice in primary care. *Journal of the American Medical Association*, *288*, 3019–3026. doi: 10.1001/jama.288.23.3019

Tamblyn, R. M., Grad, R., Gayton, D., Petrella, L., & Reid, T. (1997). Impact of inaccuracies in standardized patient portrayal and reporting on physician performance during blinded clinic visits. *Teaching and Learning in Medicine: An International Journal*, *9*, 25–38. doi: 10.1080/10401339709539809

Tamblyn, R. M., Klass, D. J., Schnabl, G. K., & Kopelow, M. L. (1991). The accuracy of standardized patient presentation. *Medical Education*, *25*, 100–109. doi: 10.1111/j.1365-2923.1991.tb00035.x

United States Medical Licensing Examination. (2013, August 1). *Step 2 CS: Scoring.* Retrieved from http://www.usmle.org/step-2-cs/#scoring

U.S. Bureau of Labor Statistics. (2013). *American time use survey: 2012 results* [Economic news release]. Retrieved from http://www.bls.gov/news.release/atus.nr0.htm

van der Vleuten, C. P. M., & Newble, D. I. (1995). How can we test clinical reasoning? *The*

*Lancet*, *345*, 1032–1034. doi: 10.1016/S0140-6736(95)90763-7

van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: state of the art. *Teaching and Learning in Medicine: An International Journal*, *2*, 58–76. doi: 10.1080/10401339009539432

van Zanten, M., Boulet, J. R., McKinley, D. W., & Whelan, G. P. (2003). Evaluating the spoken English proficiency of international medical graduates: detecting threats to the validity of standardised patient ratings. *Medical Education*, *37*, 69–76. doi: 10.1046/j.1365-2923.2003.01400.x

Vu, N. V., & Barrows, H. S. (1994). Use of standardized patients in clinical assessments: recent developments and measurement findings. *Educational Researcher*, *23*(3), 23–30. doi: 10.3102/0013189X023003023

Vu, N. V., Marcy, M. M., Colliver, J. A., Verhulst, S. J., Travis, T. A., & Barrows, H. S. (1992). Standardized (simulated) patients' accuracy in recording clinical performance check-list items. *Medical Education*, *26*, 99–104. doi: 10.1111/j.1365-2923.1992.tb00133.x

Webb, W. B., & Nolan, C. Y. (1955). Student, supervisor, and self-ratings of instructional proficiency. *Journal of Educational Psychology*, *46*, 42–46. doi: 10.1037/h0047558

West, M. A., Borrill, C., Dawson, J., Scully, J., Carter, M., Anelay, S., ... Waring, J. (2002). The link between the management of employees and patient mortality in acute hospitals. *The International Journal of Human Resource Management*, *13*, 1299–1310.

doi: 10.1080/09585190210156521

Williams, R. G. (2004). Have standardized patient examinations stood the test of time and experience? *Teaching and Learning in Medicine: An International Journal*, *16*, 215–222. doi: 10.1207/s15328015tlm1602_16