4-4-2017

Many Hands Make Light Work: Crowdsourced Ratings of Medical Student OSCE Performance

Mark Grichanik
University of South Florida, mgrichanik@gmail.com

Many Hands Make Light Work:

Crowdsourced Ratings of Medical Student OSCE Performance

by

Mark Grichanik

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

Major Professor: Michael Coovert, Ph.D.
Walter Borman, Ph.D.
Michael Brannick, Ph.D.
Chad Dubé, Ph.D.
Matthew Lineberry, Ph.D.
Joseph Vandello, Ph.D.

Date of Approval:
March 31, 2017

Keywords: OSCE, crowdsourcing, patient perspective, clinical skills, medical education

DEDICATION

This dissertation is dedicated to my family; it is their sacrifice and support that allows me to live my American Dream. To my mother, Stella Grichanik, whose dedication to her students stimulated my interest in education. To my father, Alex Grichanik, who taught me that if you're gonna do it, do it right. To my brother, Edward Grichanik, who inspires in me the courage to take the road less traveled. To my enlightened wife, Hanna Grichanik, who has gifted me the balance I didn't know I needed to savor work and life.

ACKNOWLEDGMENTS

Thank you to Michael Coovert, my major advisor, for his many years of guidance and support. Thank you for always offering me your wisdom, an open mind, and a stable platform from which I could pursue my passions.

Thank you to my committee members, Walter Borman, Michael Brannick, Chad Dubé, Matt Lineberry, and Joseph Vandello for your service. Your expertise and constructive feedback helped shape this dissertation into a meaningful work.

Thank you to the people and organizations who paved the way for me to pursue a career in medical education, Michael Brannick, Wendy Bedwell, Jan Cannon-Bowers, the Center for Advanced Medical Learning and Simulation, the Morsani College of Medicine, and the Naval Air Warfare Center – Training Systems Division.

Thank you to Matt Lineberry, a dear friend, colleague, and mentor. I always need something to look forward to and someone to look up to; you satisfy the latter in abundance. You're always courteous in saying that you'll be chasing my coattails; for now, I'm chasing yours.

Thank you to my USF classmates for many fond memories and for their unwavering support; I look forward to the privilege of a long career alongside these inspiring colleagues. Thank you to Adam Ducey, my close friend and partner through grad school; one of our projects in Selection helped germinate the idea for this dissertation. Thank you to Onursal Onen, a truly selfless friend and roommate, who helped create a home with me in Florida and who supported

me through the most trying of times. Thank you to Zaur Lachuev, a friend that keeps on giving, for always boosting my motivation to finish my “dessert station.”

Thank you to Rush Medical College for their generosity in sponsoring this research. Thank you to Keith Boyd for taking a chance and giving a budding IO psychologist an opportunity to make an impact. Thank you to the students and faculty of RMC, who inspire me to work in the service of the patients and families they heal and who make me feel righteous in my labor. Thank you to my RMC colleagues, for nurturing a creative workplace that allows me to innovate in pursuit of a more meaningful education for our future physicians. Thank you to the RMC Simulation Team, Jamie Cvengros, Ellenkate Finley, Rebecca Kravitz, and the RMC standardized patients and videographers for your tremendous effort in producing the videos for this project. Thank you to Ellenkate Finley and Ryan Horn for your incisive reviews of this manuscript.

TABLE OF CONTENTS

List of Tables iv

List of Figures vi

Abstract vii

Chapter One: Introduction 1
- Reliable, "Expert" Raters Wanted: Defining Rater Expertise and the Boundaries of Rater Reliability 5
- Beyond "Blame-the-Rater": Improving the Quality of Clinical Performance Assessment Ratings 11
- Agreeing to Disagree: Rater Divergence as Information 18
- Lay Raters: Using Non-Experts to Accomplish Expert Work 22
- Crowdsourcing: An Open Call to Novices 25
- Sorry, the Doctor is Busy: Crowdsourced Ratings in Medicine 29
- Ratee Reactions to Alternative Rating Practices 34

Chapter Two: Present Study 38
- Potential Advantages of Crowdsourced OSCE Ratings 38
- Lay Rater Accessibility to Various Rating Tasks and Formats 41
- Student Reactions to Crowdsourced OSCE Ratings 43
- Research Questions 44

Chapter Three: Study 1 Method 46
- Participants 46
  - Standardized patient raters 46
  - Faculty raters (ICS task) 47
  - Faculty raters (PE task) 47
  - Crowd raters 48
- Measures 49
- Video Stimuli 51
- Collection of Ratings 53
  - Faculty and SP ratings 53
  - Crowd ratings 54

Chapter Four: Study 1 Results 57
- Data Preparation 57
- Crowd Demographics 58
- Timing and Cost 59
- Rating Distribution and Mean Performance Levels 62

Accuracy 63
Reliability 68
Rater Reactions Questionnaire (RRX) 72
Likert-type items 72
Open-ended items 78

Chapter Five: Study 2 Method 124
Participants 124
Measures 124
Feedback Packages 125
Procedure 126

Chapter Six: Study 2 Results 129
Data Preparation 129
Rating and Feedback Quality Questionnaire (RFQQ) 129
Likert-type items 129
Forced choice package preference 132
Student Reactions Questionnaire (SRX) 133
Likert-type items 133
Open-ended items 135

Chapter Seven: Discussion 141
Who is “the Crowd”? 142
How Long Did It Take and How Much Did It Cost to Collect Crowd Ratings? 142
Were Crowd Ratings Accurate and Reliable? 145
How Did Crowd Raters Feel About the Rating Tasks? 149
Did Students Prefer Feedback from Crowd Raters? 152
How Did Students React to Crowd Ratings? 153
Limitations and Future Directions 158
Implications and Conclusion 162

References 165

Appendices 176
Appendix 1: IRB Approval Letters 177
Appendix 2: Interpersonal and Communication Skills Questionnaire (ICSF) 183
Appendix 3: Physical Exam Skills Checklist (PESC) 187
Appendix 4: Interpersonal and Communication Skills Task Crowd Rater Reactions Questionnaire (RRX-ICSF) 190
Appendix 5: Physical Exam Task Crowd Rater Reactions Questionnaire (RRX-PESC) 192
Appendix 6: Demographics Questions by Participant Type 194
Appendix 7: Sample Video Frames for ICS and PE Videos 198
Appendix 8: Sample SP Case Door Chart 199
Appendix 9: Survey Gizmo Layout for the Physical Exam Skills Task for Physician Faculty Raters 200

Appendix 10: Survey Gizmo Layout for the Interpersonal and Communication Skills Task for Behavioral Science Faculty Raters 202
Appendix 11: Survey Gizmo Layout for the Interpersonal and Communication Skills and Physical Exam Skills Tasks for Standardized Patient Raters 204
Appendix 12: Survey Gizmo Layout for the Physical Exam Skills Task for Crowd Raters 206
Appendix 13: Survey Gizmo Layout for the Interpersonal and Communication Skills Task for Crowd Raters 212
Appendix 14: Sample Amazon Mechanical Turk HIT Posting 216
Appendix 15: RRX-ICSF Content Analysis Comments by Theme 217
Appendix 16: RRX-PE Content Analysis Comments by Theme 226
Appendix 17: Rating and Feedback Quality Questionnaire for the ICS Task (RFQQ-ICS) 232
Appendix 18: Rating and Feedback Quality Questionnaire for the PE Task (RFQQ-PE) 235
Appendix 19: Student Reactions Questionnaire (SRX) 236
Appendix 20: Sample Interpersonal and Communication Skills Feedback Package Presentation 241
Appendix 21: Sample Physical Exam Skills Feedback Package Presentation 246
Appendix 22: Survey Gizmo Layout for the Student Reactions Study 248
Appendix 23: SRX Content Analysis. Advantages Comments by Theme 255
Appendix 24: SRX Content Analysis. Disadvantages Comments by Theme 259

LIST OF TABLES

Table 1:	Video and task properties	81
Table 2:	Rater schematic	82
Table 3:	Comparison of Crowd rater and U.S. demographics	83
Table 4:	Task timing (video x task x rater type)	84
Table 5:	Timing to complete rating packages (video x task x rater type)	85
Table 6:	Response distribution for ICSF checklist items (item x rater type x video)	86
Table 7:	Response distribution for ICSF global items (item x rater type x video)	87
Table 8:	Descriptive statistics for ICSF (item x rater type x video)	91
Table 9:	Response distribution for PESC (item x rater type x video)	94
Table 10:	Raw agreement for ICSF (item x rater type)	96
Table 11:	Average deviation for ICSF items (item x rater type)	97
Table 12:	Average deviation for ICSF items (rater type x skill level)	98
Table 13:	Raw agreement for PESC items (item x rater type)	99
Table 14:	Variance components from ICSF generalizability studies (item x rater type)	100
Table 15:	Phi coefficients (single rater, max raters, cost equivalent) from decision studies for ICSF items (item x rater type)	101
Table 16:	Phi coefficients (minimum raters to reliable) from decision studies for ICSF items (item x rater type)	102
Table 17:	Krippendorff's alpha coefficients for PESC items	103
Table 18:	Descriptive statistics for RRX	104
Table 19:	Theme frequencies for RRX open-ended items	105

Table 20: Descriptive statistics and mean comparisons for RFQQ 138

Table 21: Descriptive statistics and mean comparisons for SRX 139

LIST OF FIGURES

Figure 1: Mean ratings by video for ICSF global items (focus on rater type) 106

Figure 2: Mean ratings by rater type for ICSF global items (focus on video) 107

Figure 3: Mean ratings by video for ICSF checklist items (focus on rater type) 108

Figure 4: Mean ratings by rater type for ICSF checklist items (focus on video) 109

Figure 5: Decision studies for ICSF Item 1 110

Figure 6: Decision studies for ICSF Item 2 111

Figure 7: Decision studies for ICSF Item 3 112

Figure 8: Decision studies for ICSF Item 4 113

Figure 9: Decision studies for ICSF Item 5 114

Figure 10: Decision studies for ICSF Item 6 115

Figure 11: Decision studies for ICSF Item 7 116

Figure 12: Decision studies for ICSF Item 8 117

Figure 13: Decision studies for ICSF Item 9 118

Figure 14: Decision studies for ICSF Item 10 119

Figure 15: Decision studies for ICSF Item 11 120

Figure 16: Decision studies for ICSF Item 12 121

Figure 17: Decision studies for ICSF Item 13 122

Figure 18: Decision studies for ICSF Item 14 123

ABSTRACT

Clinical skills are often measured using objective structured clinical examinations (OSCEs) in healthcare professions education programs. As with assessment centers, it is challenging to provide learners with effective feedback due to burdensome human capital demands. The aim of this dissertation was to evaluate the viability of using a crowdsourced system to gather OSCE ratings and feedback. Aggregating evaluations of student performance from a crowd of patient proxies has the potential to mitigate biases associated with single-rater evaluations, allow the patient a voice as the consumer of physician behavior, improve reliability, reduce costs, improve feedback latency, and help learners develop a mental model of the diversity of patient preferences. Crowd raters, recruited through Amazon Mechanical Turk, evaluated a set of video-recorded performance episodes designed to measure interpersonal and communication (ICS) and physical exam (PE) skills. Compared to standardized patient (SP) and faculty raters, crowd raters were more lenient and less reliable, when holding the number of raters and spending constant. However, small groups of crowd raters were able to reach acceptable levels of reliability. Crowd ratings were collected within a matter of hours whereas SP and faculty ratings were returned in over 10 days. Learner reactions to crowdsourced ratings were also measured. Blind to the rater source, a majority of learners preferred the crowdsourced feedback packages over the SP and faculty packages. After learning about the potential value of crowdsourced ratings, learners were positive about crowd ratings as a complement to SP and faculty ratings, but only for evaluations of ICS (not PE) and only for formative (not summative)

applications. In particular, students valued the volume and diversity of the crowdsourced feedback and the opportunity to better understand the patient perspective. Students expressed their concerns about privacy as well as the accuracy and quality of crowd ratings. A discussion of practical implications considers future best-practices for a crowdsourced OSCE rating and feedback system.

CHAPTER ONE:

INTRODUCTION

The 21st century has ushered in many drastic changes in the way healthcare services are regulated, paid for, and delivered to the American public. In a country that has fewer physicians and poorer health outcomes than most other industrialized countries (Askin & Moore, 2014), a projected shortage of up to 90,000 physicians by the year 2025 and primary care providers, in particular, has put incredible stress on the healthcare professions education enterprise (AAMC, 2015). Additionally, healthcare systems are increasingly moving toward reimbursement models focused on the quality of patient care, safety, and satisfaction rather than the simple fee-for-service healthcare models typical of the past century (Askin & Moore, 2014).

The severe projected shortage of healthcare providers combined with a spotlight on measurable healthcare outcomes requires medical schools and residency programs to efficiently train masses of competent physicians. Modern physicians are expected to be self-directed, lifelong learners capable of delivering professional, high-quality patient care in complex systems of practice. Physicians must apply an ever-changing and increasingly complex base of medical knowledge and techniques all while maintaining excellent working relationships both with their patients as well as other healthcare providers (Englander et al., 2013).

Traditionally, models of healthcare professions education have focused on "steeping" students in the medical curriculum for a pre-determined amount of time and passing them along to the next stage of professional development (e.g., residency, independent practice) instead of

verifying that students have demonstrated proficiency in the knowledge, skills, and attitudes required of them at a particular stage of clinical practice. As Shulman summarizes, "…we should logically set levels of achievements as constants and let time act as a variable. Instead, we do exactly the opposite. We set time as a constant and have students run until their time is up. The grades we give reflect how far they have gotten in the race within the time span we have allotted (Shulman, 1970)."

In reaction to these pedagogical shortcomings, many medical education programs have begun to re-develop their curricula in accordance with competency-based medical education (CBME) models, which promote the use of an organized framework of clinical competencies to train and assess students (Frank et al., 2010; Holmboe, 2015; Holmboe et al., 2010). Kane defined clinical competence broadly as, "the degree to which an individual can use the knowledge, skills, and judgment associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice" (Kane, 1992). Bodies regulating medical education practice have also developed and medical education programs have adopted specific competency models such as the ACGME Core Competencies and Milestones, the AAMC Entrustable Professional Activities and Physician Competency Reference Set, and the Canadian Royal College's CanMEDS. These competency frameworks explicitly specify the knowledge, skills, and attitudes physicians and physician trainees need to acquire at various stages of training. In contrast to time-oriented models of medical education, CBME requires the longitudinal tracking of individual student progress against competency standards so that both students and educators can adapt the educational program to meet the developmental needs of individual learners (Holmboe et al., 2010).

While traditional medical educational practices test competence components at single moments in time, CBME specifies that competence is built incrementally and must, therefore, be assessed longitudinally. As such, learners must regularly demonstrate and educators must *directly* observe clinical competence rather than infer it based on the student's participation in the curriculum. Furthermore, CBME assessments should be conducted such that students demonstrate clinical competence not in the abstract, but rather within the clinical context.

Two popular methods of evaluating students within the clinical context are workplace-based clinical competence assessments and Objectives Structure Clinical Examinations (OSCEs) (Pangaro & McGaghie, 2015). Workplace-based assessments rely on faculty observations of trainee actions and activities within "authentic", in-vivo clinical settings. Examples of these types of assessments included case-based discussions, the mini clinical evaluation exercise (Mini-CEX), and the use of checklists and global rating scales to evaluate direct behavioral observations of trainee interactions with patients (Holmboe, Huot, Chung, Norcini, & Hawkins, 2003). The OSCE is a practical examination that aims to simulate the clinical context and evoke demonstration of the types of knowledge and skills students are required to activate in actual clinical encounters (Harden & Gleeson, 1979). This assessment method typically involves the use of actors portraying patients in a standardized, scripted role-play (i.e., standardized patients; SPs). OSCEs are popular assessment tools because clinical competence can be assessed across students within standardized clinical scenarios. OSCEs can be designed to measure a broad range of clinical competencies, such as the ability to perform a medical procedure or physical examination, communicate effectively with or counsel a patient, take a patient history, document a clinical encounter, interpret diagnostic studies, or generate a differential diagnosis. Each of these skills components can be tested within or across a variety of clinical scenarios (i.e., cases

or stations) portrayed by standardized patients. OSCEs are typically video-recorded so that students and evaluators can later review the performance episodes to generate feedback that can be used to improve performance in the examined competency domains (Stilson, 2009).

Workplace-based assessments of clinical competence and OSCEs commonly use measurement instruments that rely on the judgments of human "experts". For example, in a workplace-based assessment, clinical faculty members may be asked to directly observe a student taking a patient history in an outpatient setting and record their evaluations on a checklist or global rating form. In an OSCE, SPs may use a global rating form to evaluate a student's interpersonal and communication skills after an encounter (Regehr, Macrae, Reznick, & Szalay, 1998). However, single human ratings are notoriously unreliable due to a variety of cognitive, social and environmental biases (Williams, Klamen, & McGaghie, 2003), and multiple ratings of the same performance episode or the same clinical competence domains across multiple performance episodes are required to arrive at a stable, reliable estimate of a trainee's abilities. In addition to accurate and reliable ratings of their performance, students must receive frequent, timely, and diverse formative feedback in order to guide their development (Hodder, Rivington, Calcutt, & Hart, 1989).

Unfortunately, multiple OSCE expert ratings and feedback sessions are challenging to acquire due to logistical constraints. Faculty time is expensive, in high demand and short supply as they tend to their clinical, scholarly, and educational duties. Furthermore, since clinical faculty members or medical education program administrators are not typically trained in the psychometrics of human performance evaluations, they may view the provision of *multiple* ratings as a superfluous, and perhaps wasteful, exercise. Similarly, SP programs are resource-intensive as SPs must be trained and compensated not only to portray the case but also provide

high-quality ratings and feedback to students (Pangaro & McGaghie, 2015). Providing multiple SP ratings and feedback profiles requires additional resources such as monetary compensation for additional SP time, additional SP recruitment (if multiple raters must be present during the performance episode), or video recording and reviewing equipment (if multiple raters will provide extra ratings after the performance episode). In light of these logistical pressures, few medical education programs provide sufficient feedback or ratings.

The purpose of this dissertation is to explore a potential solution to this challenge by leveraging an emerging set of technologies, namely online crowdsourcing platforms that allow near-instant access to massive groups of proxy patients who can be used as an inexpensive, rapid, and perhaps more appropriate alternative to expert ratings and feedback in clinical competence assessments. This dissertation is organized as follows. First, I evaluate the assumptions underlying the treatment of faculty and SP raters as domain experts. Next, I explore the psychometric challenges of applying human judgments to OSCEs and describe the efficacy of various interventions that have been applied in attempts to resolve these measurement challenges. I then explore the bright side of these rating challenges by framing rater disagreement as valuable information for trainees. Next, I discuss the history of using lay people and crowds in medical and non-medical fields to perform tasks typically relegated to subject matter experts. Finally, I explore the potential benefits and challenges of using a crowdsourced system to gather OSCE evaluations from lay raters.

Reliable, "Expert" Raters Wanted: Defining Rater Expertise and the Boundaries of Rater Reliability

A traditional view of assessment holds that expert knowledge in the target domain is required for the evaluation of trainees. As Norman unequivocally states, "…there is one skill I

am prepared to recognize. I really believe that experts in a domain possess ample amounts of judgment. We have evidence all around us that people who are accomplished in an area are also able to assess reliably those who are not" (Norman, 2005). How then are experts in the domain of clinical competence assessment distinguished from non-experts, and how can we differentiate between various levels of expertise? Ericson offers this condensed history of the study of expertise:

> In the 1980s the definition of expertise based on accumulated knowledge, extensive professional experience, and peer nominations was becoming increasingly criticized. Numerous empirical examples were reported where "experts" with extensive experience and extended education were unable to make better decisions than their less skilled peers or even sometimes than their secretaries. Early studies were unable to establish superior accuracy of the peer-nominated best general physicians, when compared to a group of undistinguished physicians. Similar findings were subsequently attained for clinical psychotherapists, where more advanced training and longer professional experience were unrelated to the quality and efficiency of treatment outcomes. In addition, examinations of the cognitive mechanisms that mediated the actions of individuals exhibiting consistently superior performance revealed a complex structure that could not be accounted for by a mere accumulation of experience and knowledge (Ericsson, 2008).

If mere experience or social recognition as an expert is not the sole defining feature of expertise, what other elements are critical? A large and mature literature in the area of critical thinking considers expertise within a recognition/meta-recognition cognitive framework (Cohen,

Freeman, & Thompson, 1998) such that decision-makers first develop robust stimulus-response models of the world around them through mere exposure (recognition). For example, the clinical reasoning and medical expertise literature proposes that, through extensive experience, physicians develop a series of cognitive structures called illness scripts, which specify cue-response relationships about disease (Schmidt, Norman, & Boshuizen, 1990). Recognition is then followed by the development of meta-recognition abilities as a hallmark of expertise. Meta-recognition is the ability of the expert to monitor, mentally mark and annotate gaps, flaws, conflicts, and inconsistencies in their recognition processes (Cohen et al., 1998). Finally, another major stream of literature on expertise *per se* highlights a move away from conceptualizing expertise as simply an accumulation of knowledge or experience to a view that supports specific, complex cognitive structures as a sign of expertise. These cognitive structures, and in turn expert performance, can be improved by engaging in deliberate practice, a form of development that requires a task with a well-defined goal, and a concentrated, motivated practitioner who consistently seeks feedback, opportunities to perform, and gradually refines his or her performance (Ericsson, 2008). All of these frameworks have in common the development of an experience-dependent mental model and an ongoing active elaboration or critical evaluation of that mental model.

Another response to the criticism of expertise being defined by mere experience and social recognition is Ericsson and Smith's proposal that experts can *reliably reproduce* superior performance in a particular domain (Ericsson & Smith, 1991). If, in the domain of chess, experts consistently select the best chess move for a particular position; and in the domain of music, experts consistently play the same piece of music twice in the same manner; and in the domain of dance, experts position their bodies consistently in ideal alignment; and in the domain of

clinical decision making, experts consistently estimate more accurate pre-test probabilities, who is the expert in evaluating clinical performance encounters, and what is the yardstick of success in rating these encounters?

The unequivocal answer to the first question is that physicians are considered to be the experts in clinical performance ratings. In workplace-based clinical performance assessments – those in which students are observed performing clinical duties with actual patients – supervising clinical faculty members typically provide ratings, while in OSCEs, standardized patients typically serve as raters in addition to, or commonly in place of, faculty members. Studies of rater performance in both the industrial-organizational psychology literature as well as the medical education literature often apply the "gold-standard" model to evaluate rating accuracy. In this model, raters who have been socially nominated as knowledgeable in the domains being assessed, have extended experience in providing ratings, or have gone through a rater training program are elected as gold-standard raters. Their ratings, or an average of gold-standard ratings, then serve as a benchmark against which subsequent ratings are judged for accuracy based on absolute score deviations (Woehr & Huffcutt, 1994).

The gold-standard model is problematic because it relies on a faulty definition of expertise as mere experience in a domain. The expertise studies discussed by Ericsson (2008) above demonstrate that experience alone does not necessarily make for better performance. If raters become expert raters through deliberate practice, there is little in the medical education literature to suggest that such development regimens are used widely by medical education programs or individual practitioners serving as raters. Although several papers do address the role of deliberate practice in developing medical expertise, they focus on clinical skill acquisition rather than on skill acquisition as an accurate rater (Ericsson, 2004). Considering that almost all

Western physician competency models explicitly highlight the role of the physician as an educator, this lack of attention is unfortunate but unsurprising. With the relatively recent incorporation of "Educator" as a core competency for all physicians, physicians developing as educators have likely focused on the large repertoire of others skills necessary to be effective clinical educators. The growing zeitgeist of assessment driving learning in medical education should spur further development in this area. Finally, one could argue that frame-of-reference training incorporates a form of deliberate practice that aids in the development of expertise, and in turn, rating accuracy. Unfortunately, these training programs have shown little promise in training raters of clinical competence; an extended discussion on this topic is provided further along in this manuscript.

If we assume that the gold-standard is an elusive goal and that all "expert" raters, as they are defined in medical education (i.e., standardized patients and faculty members), are equivalent by virtue of their "expertise" in rating clinical competence, we can lean on an alternative definition of rating quality. In this second, complementary conceptualization, the metric of accomplishment is rating reliability; that is, a reliable rater, or group of raters, is a successful rater. The first question is whether raters are consistent with themselves; that is, whether an individual rater provides the same ratings when reviewing the same performance episode on two different occasions. Several studies have investigated intra-rater reliability. For example, Kalet, Earp, & Kowlowitz (1992) had faculty members re-rate medical student interviews after a 3-6 month delay and found poor intra-rater agreement in ratings of skill and behavior. However, Sturpe, Huynh, & Haines (2010) found excellent intra-rater reliability (ICC .85-.95) on checklist items in a 3-station OSCE with PharmD students after a 2-month delay. Ladyshewsky, Baker, Jones, & Nelson (2000) also found evidence of moderate intra-rater reliability in simulated

patient encounters across a variety of domains including History Taking (kappa = .63) and Physical Examination (.74) with one rater after a 2-week delay. Lang, McCord, Harvill, & Anderson (2004) found intra-rater reliability coefficients ranging between .63 and .87 after a 15-week delay using individual communication skills (e.g., rapport building, information management, agenda setting, active listening) and overall case ratings with two raters included in this portion of the study. These data suggest that raters can re-assess performance episodes with reasonable reliability, although the degree of reliability is likely dependent on the performance assessment instrument, the time delay, and the facets of performance being evaluated.

Raters diverge in their evaluations of job candidates and trainees for a wide variety of reasons, many of which are discussed in the following section. Therefore, multiple raters are typically used in assessments involving human judgment. How consistent can different raters be in evaluating the same performance episode? Williams et al. (2003) suggest that Olympic-level figure skating judges are an example of the theoretical upper limits of inter-rater reliability. For example, Weekley & Gier (1989) found that nine world-class judges were able to achieve remarkably high levels of inter-rater reliability (.93-.97). These judges watch the same exact performance, regularly receive training and re-training on clear performance criteria, focus entirely on rating, are motivated to be accurate, and most importantly, receive regular and near-instantaneous feedback on their consistency with other judges. Several of these conditions, namely regular practice, motivation to excel, and frequent performance-correcting feedback are the hallmarks of expertise development in Ericsson's expertise and deliberate practice model. Judges in clinical performance assessment rarely operate under these ideal conditions, which is one of the reasons why inter-rater reliabilities in clinical performance assessment seldom approach these theoretical upper limits.

Although a full review of the many studies in clinical performance assessment that examine inter-rater reliability is beyond the scope of this paper, several papers argue unequivocally that adding raters improves the reliability of clinical performance assessments, particularly in the assessment of non-cognitive domains such as communication skills. In a systematic review of the reliability of OSCEs, Brannick, Erol-Korkmaz, & Prewett (2011) found that adding a second rater meaningfully improves reliability both across stations (alpha 0.65 to 0.81) and across items (0.76 to 0.89). The study also found that communication items were measured less reliably across cases and more reliably within cases when compared with clinical skills (e.g., physical exam). Additionally, in a broad review examining accuracy in estimating competence, Williams et al. (2003) concluded, "The results in the studies reviewed here suggest that a minimum of 7 to 11 observer ratings will be required to estimate overall clinical competence. More ratings will be necessary to obtain stable estimates of competence in more specific competence domains (e.g., data collection, communication, responsible professional behavior), with the most ratings (up to 30) needed in the noncognitive areas." Although it should be noted that the majority of the studies in the Williams et al. (2003) review focused on workplace-based assessments rather than *structured* assessments of clinical competence as in the studies included in Brannick et al. (2011), the conclusion that multiple raters are required for clinical competence assessment is clear.

Beyond "Blame-the-Rater": Improving the Quality of Clinical Performance Assessment Ratings

The performance rating literature has identified a wide variety of cognitive, environmental, and social biases that may cause performance ratings to be inaccurate or unreliable (Williams et al., 2003). Examples of cognitive biases include difficulties with

discriminating across broad sets of performance dimensions, recalling large amounts of performance information, negative information receiving more weight than positive information, halo, central tendency, recency, primacy, contrast effects, and different raters disproportionately focusing their attention on performance information in domains that are of particular value or interest to them. Another potential cognitive bias specific to OSCEs is the cognitive demands placed on standardized patients as they have to simultaneously portray the case and keep a variety of assessment standards in mind. For example, Newlin-Canzone, Scerbo, Gliva-McConvey, & Wallace (2013) found that raters who had to act out a case and subsequently fill out a communications rating instrument missed nearly 75% of the nonverbal target behaviors portrayed by standardized students; those same participants missed fewer target behaviors (50%) when they were passively observing the encounter and not acting. Environmental biases include rater time pressure and distraction, interactions between the ratee and other factors (e.g., case complexity) or actors (e.g., other care providers, students, patients, inconsistency in standardized patient performance), the effects of the pressure to perform, inconsistent rater calibration and training, and infrequent opportunities for the rater to observe performance. Lastly, social biases include positive distortion of ratings to preserve psychological states or social relationships (e.g., the Mum Effect), rater leniency and stringency, influence from other raters (e.g., during consensus discussions), and a wide gamut of rater or ratee individual differences such as gender, race, rater motivation, rater personality, rater-ratee similarity, and liking. Together, these various biases can act as sources of measurement error in the assessment of clinical competence.

If several biases rest with the rater, there must be strategies to minimize those sources of error for which the rater is "responsible." One approach that has been commonly applied to assessments of competence to minimize rater-derived error and increase rater accuracy and

reliability is rater training. Several rater training approaches have been extensively studied in the work psychology literature, including rater error training (RET) performance dimension training (PDT), frame of reference training (FOR), and behavioral observation training (BOT) (Woehr & Huffcutt, 1994). The aim of RET is to make raters aware of common human rating error effects such as halo, leniency, severity, recency, primacy, and a disproportionate focus on negative events. A meta-analysis in the work psychology literature demonstrates that RET is effective in reducing halo and leniency and increasing rating accuracy. PDT is intended help raters avoid global judgments of ratees, and instead, to orient raters to the specific dimensions and dimension definitions on which trainee performance will be evaluated. PDT has been shown to reduce halo, increase rating accuracy, and increase leniency. FOR training helps to define performance standards within a dimension by presenting samples of behavior representative of varying levels of performance within that dimension. The training involves providing raters with practice rating opportunities and delivering feedback on how their ratings converge with a common evaluative standard within a dimension. FOR has been shown to improve rating and observational accuracy and decrease both halo and leniency. Finally, BOT focuses on how raters detect, perceive, recall, and recognize specific ratee behaviors. That is, rather than focusing on how raters evaluate behaviors, behavioral observation training focuses on how raters can accurately track behavior (e.g., taking notes, keeping a diary). BOT has been shown to improve rating and observational accuracy.

Although rater training is often used to support assessments of clinical performance in the form of faculty or standardized patient education programs (Dickter, Stielstra, & Lineberry, 2015), only a handful of studies have examined the impact of these rater training strategies as applied to medical education. Several studies using a variety of training interventions have found

no effect of rater training (Holmboe et al., 2003; Newble, Hoare, & Sheldrake, 1980; Noel et al., 1992), while others have found that training improves score reliability (Angkaw, Tran, & Haaga, 2006; Müller & Dragicevic, 2003; Müller et al., 1998). In a recent paper inspired by the Woehr & Huffcutt (1994) work, Cook, Dupras, Beckman, Thomas, & Pankratz (2009) investigated the effects of a half-day workshop focusing on PDT, BOT, RET, and especially FOR on mini-CEX ratings. The mini-CEX is a workplace-based assessment designed to evaluate clinical skills such as medical interviewing, physical examination, decision-making, and clinical reasoning. The study found no significant effects on the reliability, accuracy, or halo effects in mini-CEX scores. Cook et al. (2009) offer several explanations for why they may not have found an effect of rater training: 1) the study was underpowered given the effect size, 2) frame of reference training for a particular case or set of cases may not generalize to other cases (i.e., case specificity), 3) the workshop format was ineffective due to its relatively short duration, 4) as Williams et al. (2003) suggest, "physician raters are impervious to training", and 5) different raters observe and value different things in different domains. I come back to expand on several of these explanations further along.

If the traditional intervention, rater training, for addressing rater unreliability is ineffective, how can we develop reliable assessments of clinical competence that rely on human judgment without engaging in a perpetual campaign of "blame-the-rater"? If we acknowledge that rater unreliability is not entirely an individual difference, but instead the product of a variety of social, environmental, and cognitive factors, several practices could prove helpful. Williams et al. (2003) propose a variety of recommendations focused mainly on improving ratings of workplace-based assessments, and several of these practices apply to standardized clinical competence assessments such as OSCEs. I highlight here only those recommendations that the

present study's intervention might reasonably affect. One recommendation is the introduction of broad, systematic sampling of clinical scenarios. Because different clinical scenarios differentially activate different clinical competencies, this strategy is meant to address the limitation of evaluating broad competencies using a limited range of situations (i.e., case specificity). This strategy has been widely adopted in OSCEs through the introduction of multiple cases or stations within a single assessment, which provides opportunities for re-assessment of the domains of interest across various permutations (e.g., different chief complaints, diagnoses, tasks, patient types). A second recommendation that is applicable to OSCEs is the introduction of multiple raters to address the types of rater idiosyncrasies discussed above (e.g., leniency, severity, focus on a subset of performance domains, rater motivation). The effects of these idiosyncrasies should wash out with additional raters. Third, Williams et al. (2003) recommends that raters, faculty and SPs in the case of OSCEs, be provided with protected time to focus on making thorough and thoughtful ratings. A fourth recommendation is that clinical performance assessments for formative purposes should be separated from assessments for summative purposes such that assessments with the primary goal of supporting learning should generate timely feedback that is individualized and specific to a particular performance episode. Although the findings about the efficacy of rater training are mixed at best, a fifth recommendation is to make sure that raters are at least familiar with the assessment instrument prior to observing performance in order to focus the rater's observation behavior as well as to help the rater organize performance information.

The first two recommendations above, increasing the number of cases and raters, have been proven time and again to improve the reliability of clinical performance ratings. However, while some of the OSCE data in published studies is characteristic of typical practice in

undergraduate medical education institutions (i.e., true field studies), many studies have created special OSCE conditions to investigate various measurement properties of the OSCE method (e.g., Does adding raters or cases improve reliability?). Unfortunately, no data is readily available about the number of cases or raters medical school programs typically use for clinical performance assessments, but the Brannick et al. (2011) meta-analysis gives us a hint that these "best practices" are likely not widespread. For example, although it is unclear whether the two-rater studies come from research OSCEs or field OSCEs, only 8 of 90 across-station studies and only 6 of 43 across-item studies had a second rater. One or two raters is certainly far from the 7-11 raters rule-of-thumb Williams derived for evaluating overall clinical competence and even further from the up-to-30 ratings for specific, non-cognitive competencies. However, it is important to remember that the Williams et al. (2003) numbers were based on studies of workplace-based, rather than standardized, clinical performance and should, therefore, be adjusted downward in consideration of the more constrained clinical environment and stimuli found in OSCEs.

Although securing more raters and protecting time expressly for the purpose of assessment is now mantra in the clinical performance assessment community, the situation on the ground is challenging because of logistical and resource constraints. Standardized patients, typically paid $17-35/hour (Sun, n.d.), quickly become expensive as programs add multiple raters. Faculty time is even more expensive than SP time, and educational responsibilities often compete for time with clinical duties. Stretched faculty members, if available, can take a long time to review performance episodes and even longer to deliver feedback to students.

With respect to the merits of adding cases/stations to improve reliability, Govaerts, Van der Vleuten, & Schuwirth (2002) present generalizability coefficients from a midwifery OSCE as

a function of the number of cases and raters. For example, for six cases the generalizability coefficients are .53 and .58 for one and two raters, respectively. For 20 cases, the generalizability coefficients are .79 and .82 for one and two raters, respectively. Brannick et al. (2011), in a meta-analysis of OSCE reliability, found more appreciable gains for adding a second rater (.65 to .81 for across-station estimates; .76 to .89 for across-item estimates). Govaerts et al. (2002) conclude that increasing the number of stations gets examiners to higher levels of reliability more effectively than increasing the number of raters per case. Although this may be true, Govaerts et al. (2002) fail to consider the tradeoff in resources required for adding cases versus adding raters. The addition of a case requires significantly more resources than the addition of a rater. Adding a case requires case development and piloting, booking more simulation center time, additional time for standardized patient training, additional SP portrayal time, additional SPs (if finishing the OSCE within a particular time frame is important), additional administrator time, extended student time away from other curricular activities, and additional video storage space (if recording). Assuming that SPs are the only raters, the addition of raters only requires video-recording of the performance episodes (something most programs already do), computers to replay the videos, and additional SP time to review the videos and provide second, third, fourth, etc ratings. If the same SPs are used for off-line rating as those who portrayed the case, no additional case training is required. However, it is important to note that merely adding raters to a narrow range of cases is not sufficient for medical education applications; we expect students to have broad case exposure and to demonstrate a variety of facets of their clinical competence across a variety of clinical scenarios. A balance between the two approaches is likely to be more logistically feasible and acceptable to clinical educators.

Finally, it is important to note that OSCEs are structurally similar to assessment centers in the work psychology literature (Arthur & Day, 2011). Both techniques ask participants to perform job-relevant tasks across several cases or stations, and human raters or assessors evaluate participant performance across several competencies or domains. OSCEs and assessment centers also share common measurement challenges. For example, large portions of variance are often attributed to exercise/method factors or cases/stations when much of the variance is desired to come from competencies or performance dimensions across stations. As such, many of the best-practices in clinical performance assessments are consistent with recommendations for assessments centers. These include decreasing the number of competencies/domains while increasing the number of cases/stations, increasing the number of raters/assessors to improve reliability, and providing raters/assessors with training.

Agreeing to Disagree: Rater Divergence as Information

Some researchers have suggested that we shift our attention away from developing techniques that attempt to improve, per se, the accuracy of ratings through either instrument format refinement (DeNisi & Sonesh, 2011; Landy & Farr, 1980) or rater training (Cook et al., 2009), and concentrate instead on broader changes to performance assessment systems. One such focus has been on the provision of assessment feedback to improve performance. Although feedback is a large research area with a wide array of topics such as feedback delivery and feedback acceptance, I focus here on one facet of feedback of particular import to the present study, feedback diversity. Cook and colleagues summarize the value of feedback diversity clearly in their reflections about the frame-of-reference calibration exercise during a rater training workshop, "[Raters] often disagreed…and even prolonged discussion did not guarantee consensus…while inconsistent ratings will lower inter-rater reliability and accuracy, differences

of opinion among [raters] may actually enhance formative feedback by illustrating alternative approaches to specific tasks" (Cook et al., 2009).

The notion that rater disagreement can be viewed as information, rather than simply measurement error, motivated the development of 360-degree feedback systems in organizations. The idea is that different sources (e.g., customers, subordinates, colleagues, supervisors, vendors, faculty, student peers, self) observe different facets or have different interpretations of a target's behavior, and that aggregating feedback from these other various sources provides a more complete picture of an employee or student's performance. The employee or student is then supposed to use this information to develop action plans for performance improvement (Brett & Atwater, 2001).

Within clinical performance ratings, it is clear per the discussion above that human raters, within and across sources, reliably disagree. If the disagreement is expected, and if that disagreement is meaningful, perhaps we can improve the manner in which students use OSCE feedback to change behavior by increasing the number of sources as well as the number of ratings within each source. Such a system might include ratings by multiple faculty members, student peers, interprofessional student peers (e.g., nursing, physician assistant, and occupational therapy students), and standardized patients. If the most important outcomes for a physician are excellent patient care and high patient satisfaction, then patients should also be included as an important feedback source. The idea of including the client as a critical input in assessment development and scoring has been around for some time. For example, (Weekley & Jones, 1997) advocated the development of a rational scoring key driven by customer preferences for desirable behaviors. Indeed, Weekley & Jones (1997) found that situational judgment tests scored using a client-derived key were predictive of job performance.

Similarly, patient perceptions of physician competence should be included directly early in medical student training because ratings of student clinical competence by expert evaluators such as SPs and faculty during training might not reliably predict patient perceptions of effectiveness. For example, Mazor, Ockene, Rogers, Carlin, & Quirk (2005) conducted a study to evaluate whether lay people perceived students who scored better on communication checklists as more effective communicators. Trained raters (graduate students in psychology) used a checklist composed of patient communication behaviors considered effective by clinical faculty to evaluate student-SP interactions during an OSCE. Lay volunteers, acting as analogue patients, then watched these same videos and completed the American Board of Internal Medicine patient satisfaction questionnaire, a scale heavily loaded with communication items and designed to be used by patients with no training. Despite the fact that greater than 50% of patient proxies rated 18 of the 23 items on the checklist as "Very Important" for effective communication, there was no relationship between the checklist scores and the patient satisfaction measure in four of five cases. Further analyses demonstrated that particular checklist items were differentially discriminating in patient satisfaction scores across cases, with only two items, "Presented information clearly" and "Moved through the encounter in an efficient manner," presenting consistently in the top five list of discriminating items within each case. Mazor et al. (2005) provide two potential explanations of why they found little correspondence between the checklist scores and the patient satisfaction scores: 1) effective communication behaviors are likely dependent on the case and clinical context, and 2) the checklist was deficient in representing the behaviors critical for patient perceptions of communication efficacy (e.g., behaviors that "turn patients off" such as inappropriate tone of voice).

In order to elucidate the types of behaviors that various stakeholders involved in clinical competence assessment value, Mazor et al. (2007) conducted a follow-up study with three physicians, three SPs, and three lay raters acting as patient proxies. As the study participants were rating 20 videos from four OSCE stations on professionalism, they were asked to think aloud. Their comments were recorded and coded in order to understand how different types of raters reason through clinical performance rating assignments. In analyzing the rater comments elicited through the think-aloud protocol, Mazor et al. (2007) found different rater types made comments about different dimensions (e.g., physicians and SP commented more on the introduction, whereas lay raters commented more on explanation). However, raters within a rater type also diverged significantly in the comments. The comments also revealed at least one discrepancy in 19 of 20 encounters. Raters watching the same exact video either didn't agree whether a behavior even occurred or had diametrically opposed evaluations of the same behaviors (e.g., a positive vs. negative evaluative comment about a particular behavior). In light of these findings, Mazor et al. (2007) concludes, "…recommendations to focus evaluation efforts on behaviors may be more difficult to implement than anticipated. Efforts to identify and define professional and unprofessional behaviors may require input not only from multiple constituencies, but also from multiple representatives within each group." Moreover, this data highlights the richness and diversity different raters bring to clinical performance assessments and, as a corollary, that the same physician behaviors may evoke different reactions in different patients. This research powerfully demonstrates Ilgen, Ma, Hatala, & Cook (2015) and Cook et al.'s (2009) assertions that different raters, within and across rater types, observe and value different things across various domains and that raters are likely to disagree, despite training and common assessment tools, about the definitions of "appropriate or adequate" performance.

Lay Raters: Using Non-Experts to Accomplish Expert Work

Given that it is resource-intensive for expert raters to provide individual ratings of performance, several studies have investigated the possibility of using lay raters in place of experts. Provided that lay ratings can approximate the accuracy and reliability levels of expert raters, using lay raters can be an economical alternative.

For that example, Shohamy, Gordon, & Kraemer (1992) found lay raters are capable of reliably rating writing samples. In this study, four groups of five raters were asked to evaluate 50 writing samples using several scales. The first group was composed of professional English teachers who received training, the second of professional English teachers who did not receive training, the third of lay native English speakers who received training, and the fourth of lay native English speakers who did not receive training. Trained raters were oriented to the performance scales and participated in FOR training. Shohamy et al. (1992) found that there was no overall difference in the reliability coefficients between the professional English teachers and the native English speaking lay raters. While all the reliability coefficients were reasonably high (.80-.93), trained raters were slightly more reliable that untrained raters, but there was no main effect of expertise.

In another study, Dickter et al. (2015) examined SP and non-SP (graduate student) ratings of interprofessional practice scenarios using a TeamSTEPPS-derived instrument that focused on team structure, leadership, and communication. Both groups were trained in an 8-hour workshop, which included a review of the instrument, frame-of-reference training, and discussion of domain-relevant behaviors with video examples. Dickter et al. (2015) found that a 6-rater panel of SPs yielded a generalizability coefficient of 0.95 while the non-SP panel demonstrated a .80

coefficient. However, the D-study showed that a *single* SP rater yielded a lower coefficient of .74, and a single non-SP rater was well below acceptable reliability levels at .40.

Zanetti et al. (2010) asked three experienced SPs, three physicians, and three lay raters to evaluate professionalism across four cases. Raters were instructed to view performance videos as though they were the patient (i.e., the SP in the encounter). Raters were provided limited formal training; they were only given a practice video to review in conjunction with the rating instrument, but no explanation of the instrument was provided. In the G study, Zanetti et al. (2010) found that 20% and 15% of the variability was attributable to the student in the physician and lay rater conditions, respectively. Only 5% of the variability was attributed to students in the SP condition. They also found that the G coefficients were highest for the physician raters (.75), followed by the lay raters (.68), and SP raters were the lowest (.53). In the subsequent D study, they projected it would take 4 physician raters, 13 SP raters, and 7 lay raters to reach acceptable reliability levels of .80.

Motivated by resource constraints, Berger et al. (2012) had lay raters use a comprehensive scoring guide to grade student documentation (i.e., notes) generated from an OSCE to determine whether lay raters could reliably assess clinical reasoning. An 8-station OSCE with 180 students generates 1,440 notes. If each note takes five minutes to score, this constitutes 120 hours of work for expert raters – a pricey proposition. Berger et al. (2012) employed an analytic, rather than global, scoring technique such that raters used a detailed rubric to score component parts that were later summed to yield a global score. Kappas for the component scores (supporting evidence, evidence against, diagnostic workup) between the physician rater and the lay rater ranged between .62 and .69 while the summed global scores, which introduced additional criteria (e.g., students were deducted a point if they didn't mention a

particular type of diagnosis), had a kappa of .56. The authors conclude that, with adequate training (they do not describe their training regimen) and a detailed rubric, lay raters can score notes in alignment with faculty raters.

Multiple studies have also investigated the use of medical students as OSCE raters, citing the same resource constraints described in other studies investigating lay raters. While most schools have a large contingent of senior medical students who can serve as inexpensive peer raters (e.g., $20/hour work study), the opportunity cost of pulling students away from their own curriculum is tremendous. Furthermore, experienced medical students are not pure lay raters because they have experience in the types of clinical competencies assessed in OSCEs. In one study, Chenot et al. (2007) compared physician ratings with student tutor ratings in a 4-station OSCE. Inter-rater agreement (kappa) ranged from .41 to .64. In a second study, Moineau, Power, Pion, Wood, & Humphrey-Murto (2010) found similar results in comparing faculty and students OSCE ratings, with correlations between the two rater types ranging from .56 to .86 on checklist scores for individual stations and the correlation between rater types ranging from .23 to .78 for global rating scores across individual stations.

Collectively, this set of studies demonstrates that lay raters are generally able to reliably assess performance as well as "expert" raters across a variety of dimensions. However, one might need to employ a greater number of lay raters to approximate the reliability of fewer experts. Unfortunately, few of these studies discuss the feasibility of scaling lay ratings. If lay raters are to be used consistently in assessments of clinical competence, a large pool of raters must be recruited, provided with access to performance episodes and rating instruments, and paid. Furthermore, lay raters must be able to provide students with feedback within the same period of time or faster than SPs or faculty assessors can provide. The following section

discusses crowdsourcing, a technique which may offer reliable access to large groups of lay raters.

Crowdsourcing: An Open Call to Novices

The advent of crowdsourcing platforms offers an enticing alternative to address some of the logistical, resource, and bias challenges associated with rating clinical performance assessments. This technique is appealing for organizations seeking "better, cheaper, faster" assessment solutions delivered on digital platforms. Below, I define crowdsourcing and survey several current applications of crowdsourcing technology across a variety of disciplines, including in medical education.

Many parties have proposed definitions of crowdsourcing, a moniker still in its infancy. Estellés-Arolas (2012) systematically aggregated this wide array of interpretations of the term "crowdsourcing" and developed a taxonomy of differentiating characteristics that consists of three higher order categories: About the crowd (who forms it, what is has to do, what it gets in return), About the initiator (who it is, what they get in return for the work of the crowd), and About the process (the type of process it is, the type of call used, the medium used). Based on this structure and an investigation of a wide variety of crowdsourcing applications, Estellés-Arolas (2012) defined crowdsourcing as follows:

> "Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always

> entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken."

Threadless, for example, is an online company that gathers designs from artists around the world and prints them on apparel (e.g., t-shirts) bought by Threadless customers. Threadless is a clear example of crowdsourcing per the Estellés-Arolas definition, because the crowd can be readily identified (artists from all over the world), it has a task (create graphic designs), the crowd is compensated (recognition as an artist and profit-sharing in merchandise sold with their designs), there is a crowdsourcer (the company Threadless), the crowdsourcer benefits (Threadless sales), there is well-defined process (submitting designs and Threadless users rating them), and there is an open call (using the Threadless website) through the Internet. I will return to components of this definition as I propose the use of crowdsourcing to obtain clinical performance assessment ratings to ensure that this proposition fits with the definition of crowdsourcing.

The idea of aggregating individual contributions from a crowd to accomplish more than could otherwise be accomplished by any individual member of the crowd or of the individuals working together directly was popularized by James Surowiecki. In The Wisdom of Crowds, Surowiecki (2005) opens with Francis Galton's now-famous early demonstration of crowdsourcing. Galton aggregated approximately 800 individual entries from villagers competing at a livestock fair to guess an ox's weight. Calculating the central tendency of the individual entries resulted in the most accurate estimate of the ox's actual weight (Galton, 1907).

Many applications of crowdsourcing have demonstrated that solutions generated by aggregations of novices can converge with, and sometimes best, those produced by high-priced experts. For more than a decade, scientists were unable to solve the structure of a retrovirus enzyme, which was a critical link in developing antiretroviral drugs for HIV/AIDS. Stumped, scientists at the University of Washington created Foldit, an online protein-folding game, to challenge thousands of gamers from all over the world to come up with a solution; the gamers managed to model the enzyme accurately in just three weeks (Khatib et al., 2011). This research demonstrates that crowdsourcing can "channel human intuition" and "turn novices into domain experts capable of producing first-class scientific discoveries." In another example of crowdsourcing, PlateMate technology allows a user to snap a photograph of a plate of food and upload it to a crowd for caloric evaluation. The crowd's caloric estimates prove to be as accurate as those of costly expert nutritionists and superior to those generated by computational algorithms (Noronha, Hysen, Zhang, & Gajos, 2011). As one of PlateMates designers put it, "We can take things that used to require experts and do them with crowds." Turner-McGrievy, Helander, Kaipainen, Perez-Macias, & Korhonen (2014) similarly found that novice peers were able to judge the healthiness of plates of food as well as raters formally trained to evaluate meals using specific dietary guidelines.

Crowdsourcing has also been applied to developing assessments used in organizations. For example, Legree, Psotka, Tremble, & Bourne (2005) used a consensus-based measurement system to assess emotional intelligence and to score situational judgment tests (SJTs). In developing SJTs, experts are typically called upon to 1) identify and describe situations, 2) determine appropriate interpretations and responses, and 3) develop scoring criteria to assess performance. This study found that the large populations of non-experts were able to develop

scoring standards that converged with those of experts. In another recent example, the Behavioural Insights Team (BIT), backed by the British government, tasked a crowd with evaluating hypothetical job candidates on their responses to a behavioral-based interview question (Michel, 2016). The BIT found that relatively small crowd panels (7 raters) were capable of reliably rank-ordering candidates, even when the differences in performance among candidates were small. The BIT has since developed the Applied platform to apply their findings to real selection systems. The Applied platform anonymizes applications by removing individual identifiers (e.g., age, gender, race) that can bias evaluations of candidates, reduces evaluator cognitive load by presenting responses from all candidates for a single item at a time, and aggregating ratings from multiple internal evaluators ("Applied," n.d.).

Because humans can reliably perform tasks that computers currently cannot (e.g., identifying distorted images, coding facial affect), crowdsourcing has also been extensively applied as an inexpensive, scalable solution to annotate large amounts of photo and video data. For example, Borsboom (2012) created a variant of the popular "Guess Who" game to enable crowd raters to identify human facial affect using the Facial Action Coding System (FACS), a coding scheme that typically requires raters to participate in nearly 100 hours of training and many more hours of practice. In this game, an array of faces is displayed to two players. As in "Guess Who," player one must identify the photo that player two has chosen by asking a series of questions. Player one asks questions of player two about specific facial action units (e.g., "Does your picked card have a raised eyebrow?") that might be present in the selected photo. These responses, aggregated across a crowd, inform the action units present within the particular image; the presence and absence of particular action units, in turn, serves to identify the emotion present in the photo (e.g., happiness is the composition of two action units: raised cheek and lip

corner pulled). Borsboom (2012) found that the crowd was able to code action units and identify emotions to accurately arrive at the ground truth as well as expert FACs coders.

As a final example, the Pentagon and the Office of the Director of National Intelligence are testing the Aggregative Contingent Estimation System (ACES), a crowdsourcing solution with 1800 participants used for soliciting predictions about national security, global conflict, and social and economic futures. The military and intelligence communities are interested in whether such a technique can lead to better solutions than the aggregation of professional opinions, which is often subject to biases. The system was able to predict, with 66% accuracy, the repeal of Don't Ask, Don't Tell and the step-down of Yemeni President Ali Abdullah Saleh with 80% accuracy (Parsons, 2011).

Sorry, the Doctor is Busy: Crowdsourced Ratings in Medicine

The discipline of surgical education has one of the most active and well-established research communities in the medical education field around performance assessment. This is likely because surgical specialties require the development, certification, and maintenance of a vast array of procedural skills, which acutely determine patient outcomes. The surgical education literature has explored a wide array of training and assessment approaches to provide trainees with sufficient practice and developmental feedback opportunities. Two approaches, in particular, simulation and global surgical performance rating scales, have shown promise but are especially resource-intensive as busy surgeon raters must pull away from clinical practice and teaching to review real-time or video-recorded performance episodes. As such, the field has sought more affordable, yet "valid," avenues to capture and evaluate performance information. One of these techniques has been the development of simulator performance metrics, another has focused on automated video analysis algorithms to extract surgical performance information

(e.g., how efficiently the surgeons manipulates laparoscopic instruments), and an entirely new approach has employed crowds to rate videos of surgical performance episodes.

A single, yet fascinating, line of research on crowdsourced surgical ratings exists in the extant literature; I expand on these studies in detail because, to my knowledge, they are the only investigations of crowdsourcing to gather ratings of clinical competence and because the particulars of the study designs are pertinent to the present study.

Chen et al. (2014) was the first study in this line. A group of ten experienced robotic surgeons established a gold-standard, or ground truth, by rating a two-minute robotic surgery knot-tying video using an adapted version of GEARS, a global rating form intended to evaluate skills in robotic surgery. 611 crowd raters, recruited from Facebook and Amazon's Mechanical Turk (MTurk), were then qualified into the study using a forced-choice task wherein two videos of a surgical task, one demonstrating high skill and the other, intermediate skill, were presented side-by-side. Crowd raters were advanced into the main study if they were able to discriminate between high and low performance in this single set of videos. The 476 qualified raters subsequently watched the target video and provided ratings on three dimensions (depth perception, bimanual dexterity, and efficiency). The 95% confidence intervals around the mean score across the three dimensions for both the MTurk and Facebook crowd raters fell entirely within the pre-determined range of +/- 1 of the gold-standard mean. Chen et al. (2014) also found that this procedure produced ratings quickly. Whereas it took 24 days to gather responses from the expert surgeons, it took 24 hours to gather 409 responses on MTurk. Although the crowd raters converged on the ground truth aggregated mean, no information is presented about crowd rater performance across the individual domains (e.g., inter-domain correlations, domain means). Presumably, raters could be more or less accurate depending on the domain being

evaluated. Furthermore, no information is available about the reliability of the crowd ratings since only one video was used. Each MTurk rater was also paid $1 for providing a rating, which equals $409 for a set of ratings for a single 2-minute video. If one of the arguments for using crowdsourcing is to mitigate resource constraints, this approach certainly did not accomplish that, even by surgeon salary standards. A more appropriate approach would have been to collect a limited amount of ratings initially and to continue sampling until the a priori criterion was reached (95% crowd mean CI contained within +/- 1 of the mean expert ratings). Indeed, even in this study, the 67 uncompensated Facebook raters (far fewer than the MTurk workers) met the a priori criteria at a cost of $0. Finally, without any control condition (or variance information) it is possible that any set of raters would have converged on the same solution by chance. If the design had permitted the screened-out raters (removed because they were not able to discriminate between levels of surgical skill) to participate in the main task, that set of raters would have served as a natural control condition such that we might expect that set of raters not to converge with the expert ratings as well as the screened-in raters did.

In a follow-up study, Holst, Kowalewski, White, Brand, Harper, Sorensen, Kirsch, et al. (2015a) introduced five videos across the performance continuum, ranging from junior residents to senior attendings) to deal with the single video limitation in the Chen et al. (2014) study. Holst et al. (2015a) also limited the number of crowd ratings collected based on crowd-size simulations of the Chen et al. (2014) data. Both the mean crowd ratings and the mean expert ratings placed the five videos in the same rank order of performance. The inter-rater reliability across judges, reported as Cronbach's alpha, was high at .91. It took on average 2 hours and 50 minutes to gather 50 crowd worker ratings. In a second task, crowd workers watched a video of a resident performing a surgical task "cold," another video of that same resident performing the task after

warming up ("warm"), and a third video of an experienced surgeon performing the task. The crowd ratings were again able to discriminate between the three performance episodes in the predicted direction, such that the expert video received the highest performance ratings, followed by the warm video, and then the cold video. The important contribution of this study is that crowd raters are able to discriminate among varying levels of performance in highly technical tasks.

In the third study in this line of surgical crowdsourcing, Holst, Kowalewski, White, Brand, Harper, Sorensen, Truong, et al. (2015b) introduced videos of surgical tasks performed on live tissue. The recorded performance episodes in the previous two studies were of Fundamentals of Laparoscopic Surgery "dry-lab" tasks wherein surgeons use real surgical instruments to perform tasks that simulate the types of movements they are required to perform in real tissue. For example, in the peg transfer task surgeons must quickly move small plastic objects from one pegboard to another by transferring the plastic object from the grasper tool in one hand to the grasper tool in the other hand; this is meant to simulate, among other actions, the handoff of a needle from one grasper to another in a live suturing task. Although the dry-lab tasks are representative of the psychomotor skills required in live surgeries, Holst et al. (2015b) argued that live tissue introduces a significant degree of variability and crowd raters might not be able to discriminate performance in live tissue tasks as well as they can in dry-lab tasks. In this study, Holst et al. (2015b) largely replicated Holst et al. (2015a), albeit with videos of surgical tasks performed on live porcine tissue. It took approximately 4.5 hours to collect crowd ratings for 12 videos (50/video) whereas it took two weeks to gather ratings from all 7 experts. Inter-rater reliability was high (alpha = .93), and the correlation between the mean expert and crowd scores

was .95, which demonstrates that crowd raters are able to rate live tissue tasks as reliably as dry-lab tasks.

A fourth study by White et al. (2015) had experts and crowd raters evaluate two separate dry-lab tasks, rocking pegboard and suturing. This study notably included more videos that the previous investigations, 49 for each task across a variety of skill levels. Three surgeons and 30 crowd workers rated each video. One of the tasks was compensated at 25 cents per rating, while the second task was compensated at 50 cents per rating; White et al. found that ratings were returned significantly faster at the higher compensation point (108 hours total for 25 cents; 9 hours total for 50 cents). The reliability of mean ratings between the crowd and experts was 0.86 (Cronbach's alpha) for the rocking pegboard task and .92 for the suturing task. Surgeon scores and crowd scores were correlated .79 and .86 for the rocking pegboard and suturing tasks, respectively. White et al. (2015) also present an interesting analysis of rating costs. Acquiring 30 ratings at 50 cents each plus 10% overhead to the MTurk costs $16.50 per video rated while asking three surgeons to provide a rating (5-10 minutes per video) costs $54-$108 per video rated when one considers a surgeon's annual salary and benefits.

Finally, a fifth study by Aghdasi et al. (2015) presented 26 videos of cricothyrotomy procedures to crowd raters who were screened using the same forced-choice procedure as in the Chen et al. (2014) and Holst et al. (2015a, 2015b) studies. Raters used the OSATS instrument, which consists of six 5-point, behaviorally anchored questions: respect for tissue, time and motion, instrument handling, knowledge of instruments, flow of operation, and knowledge of the procedure. Three expert surgeons also evaluated each video. Thirty crowd ratings per video were collected within 10 hours, while it took 60 days for the expert surgeons to rate the videos. 85% of crowd-sourced average scores fell within 5 points of the expert surgeons' scores, and the two

score sets were positively related (r = .83). Cronbach's alpha, used as a measure of inter-rater reliability, was .83. Finally, participants were classified into one of three skills groups based on their OSATs scores (>20 = skilled, 10-20 = average, <10 = beginner), and skill classifications were consistent across the expert and crowd conditions with the exception of 2 of 26 cases.

This set of crowdsourcing studies is consistent with the lay rater studies described in the previous section in that absolute novices are able to reliably accomplish rating tasks typically relegated to domain experts. Crowdsourcing in surgery is particularly fascinating because it shows that crowd raters can converge on expert ratings in a highly technical domain to which presumably few crowd raters have ever had exposure, let alone rating responsibility for.

Ratee Reactions to Alternative Rating Practices

In addition to using performance assessment scores to ensure competence, evaluate the curriculum, and predict future performance, the major goal for performance assessments in education settings is to motivate student behavioral change. Formative assessments should provide students with feedback they can use to improve competence across a variety of domains. Although the measurement properties (i.e., accuracy and reliability) of the performance assessment system are important, other factors such as student reactions to and satisfaction with the performance assessment and process may determine whether students "buy in" and actually use the feedback generated by the assessment to improve performance. As DeNisi & Sonesh (2011) says about workplace performance appraisal systems, "rating accuracy, per se, should not be the sole focus of research to improve ratings. Rating accuracy is important, but it is equally important that employees perceive the ratings as fair and accurate. Only when employees perceive the appraisal processes as fair and are satisfied with them, will they be motivated to improve their performance."

The critical importance of examinee and employee reactions is highlighted in several research streams, including employee selection, training, and performance appraisal. Within the context of employee selection, research shows that job candidates differentially prefer selection tools based on factors such as opportunity to perform, face validity, justice, and job relatedness as well as processes embedded within the job selection cycle such as consistency, feedback, honesty, and two-way communication (N. Anderson, Salgado, & Hülsheger, 2010; Hausknecht, Day, & Thomas, 2004). Furthermore, perceptions of the selection procedures influence important outcomes such as test motivation, performance on selection batteries, organizational attractiveness, recommendation intentions, offer acceptance intentions, litigation intentions, and product purchase intentions (Hausknecht et al., 2004). Reactions are also important with respect to training. For example, trainee affect and utility reactions about a particular training program predict training transfer and learning (Alliger, Tamnenbaum, Bennett, Traver, & Shotland, 1998; Blume, Ford, & Baldwin, 2010). Reactions to training also predict post-training motivation and post-training self-efficacy, both essential components for self-improvement (Sitzmann, Brown, Casper, Ely, & Zimmerman, 2008). Employee satisfaction in performance appraisal systems also predicts intentions to leave the organization, job performance, and job commitment (Ellickson & Logsdon, 2002; Horvath & Andrews, 2007).

Multiple studies of non-physician raters evaluating physician or medical student performance have measured reactions. Wenrich, Carline, Giles, & Ramsey (1993) study of nurses rating physicians found that 80% of physicians provided a supportive rating of the adequacy of using nurses to evaluate physician communication skills and humanistic qualities, although physicians were less enthusiastic (46% supportive ratings) about nurses rating them on clinical skills. Chenot et al. (2007) reported that 64% of examinees perceived that there would be

no difference in the OSCE grade when measured by student or faculty examiners (27% thought it would be higher, 8% thought it would be lower). 92% of the examinees also believed that student raters would be as objective as faculty and 85% felt that the student raters had assessed them accurately. In another study of senior students providing OSCE rating, Moineau et al. (2010) found that examinees were comfortable being evaluated by student examiners (SE) on history-taking, communication, and physical exam skills (M = 4.68-4.68/5), that SE OSCE ratings would be acceptable for formative (4.61) and summative (3.98) purposes, that the SE feedback was constructive (4.73) and delivered in an appropriate manner (4.88), the SE OSCEs are worthwhile (4.43), and disagreed that there was counterproductive tension between themselves and a peer examiner (1.89). Despite these positive evaluations of student examiners, examinees were still mixed on whether they preferred faculty to student examiners (2.86).

Several studies have also investigated student reactions to various performance assessments and processes in medical education. Duffield & Spencer (2002) measured the "acceptability" of assessments in a single medical school. Nearly 90% or more of the students agreed that "in an ideal world," assessments should be used to ensure competence, provide feedback, guide student learning, and evaluate the curriculum, whereas only about half of student agreed that assessments should be used to predict performance as a doctor. Particular forms of assessment were seen as differentially "fair" with data interpretation papers the most fair (89% agreement) and clinical rotation ratings the least fair (38% agreement). 79% of students agreed that OSCEs were fair assessments, although student comments revealed several reasons why OSCEs might be perceived as unfair such as exaggerated expectations for the allotted encounter time, the artificiality of the situation, and the stress associated with performing in an OSCE.

Finally, although a wide variety of system features determine reactions to and satisfaction with training, selection, and appraisal systems. I present here only the determinants of satisfaction with appraisal systems because the employee appraisal process aligns most closely with the clinical performance assessment process. DeNisi & Sonesh (2011) summarizes the following determinants of satisfaction with employee appraisal systems: procedural, distributive, and interactional justice perceptions, two-way communication, timeliness of appraisals, frequent and consistent feedback, trust towards the evaluator, perceptions that the performance appraisal fosters future improvement, identification of individual strengths and weaknesses, and involvement in the process of developing performance appraisals. I return to these criteria in the following section as I discuss the viability of crowdsourcing OSCE ratings.

CHAPTER TWO:

PRESENT STUDY

If lay gamers are able to solve complex protein-modeling problems, and if crowds can accurately estimate the calories on a plate of food, and if lay raters can converge with surgeon raters in evaluating complex surgical tasks, then perhaps lay raters may also be able to reliably rate clinical competence in OSCE performance episodes. As such, this study proposes the development of a crowdsourced OSCE rating system to complement faculty and SP ratings. In the following section, I discuss potential advantages, lay rater format considerations, and the importance of reactions in a crowdsourced OSCE system.

Potential Advantages of Crowdsourced OSCE Ratings

There are many potential logistical and psychometric advantages of using a crowdsourced rater system in OSCE applications. First, multiple raters are required to reliably estimate global clinical competence, even more ratings are required to assess individual performance domains, and yet more are necessary to estimate performance in important non-cognitive domains such as communication and interpersonal skills (Brannick et al., 2011; Williams et al., 2003). Collecting multiple SP and clinical faculty ratings is resource-intensive. Faculty and SP ratings are expensive to collect when one considers direct costs such as faculty and SP time and indirect opportunity costs such as SP effort being re-directed away from mastering the portrayal of a case or faculty effort being diverted away from teaching, mentoring, or clinical practice. Depending on the volume of lay raters required for a stable measurement and

the compensation per performance episode reviewed, crowdsourced lay ratings may offer significant cost savings over using SP or faculty raters. Although more lay raters than faculty or SP raters may be required per performance episode, these costs are likely to be offset by reduced costs in compensation, training, and management.

Second, delivering feedback quickly and frequently should be a primary focus of assessment systems designed to motivate student behavioral change, but "despite its importance in professional development, feedback is insubstantial and infrequently given" (Cook et al., 2009). Thus, rapid evaluation of performance episodes is critical. The surgical crowdsourcing studies demonstrate definitively that many more lay ratings can be gathered much faster than a significantly smaller number of physician ratings. If performance episodes are digitized through audio or video recording and posted as rating tasks on a crowdsourcing platform, educators can increase the throughput of reviews of student performance because a nearly limitless pool of assessors can work on the rating tasks in a parallel fashion. Additionally, ratings can be completed based on the urgency with which the educator needs the assessment results. For example, if the educator requires a faster turnaround on the assessments, they can increase the monetary compensation for each rating in order to increase interest and recruit more crowd raters faster. Also, although crowdsourcing clinical performance ratings does not offer a direct solution for the case-specificity challenge, educators may find that they are able to add additional cases to OSCEs if the rate-limiting process of securing ratings and providing feedback is improved by the additional bandwidth possible with a crowdsourced system.

Third, the psychometric properties of the clinical performance ratings should improve with a crowdsourced assessment system. Because multiple assessors can evaluate a single performance episode, reliability should be equivalent or improved over the fewer number of

faculty or SP ratings typically available. Additionally, the effects of biases such as liking, similarity, rater-ratee interactions (e.g., common race), leniency, or stringency that may be present in a single rating are more likely to wash out across a larger group of ratings. Furthermore, if ratings are collected in a piecemeal fashion (e.g., separate raters for physical exam and communication) and the component scores are later aggregated, this may decrease other psychometric errors, such as halo. That is, generalized impressions formed as a result of poor performance in one scenario or on one task should not affect ratings on another one. Sackett & Wilson (1982), in fact, recommended the development of such a "disassembled assessment center" where ratings on exercises are made independently and then mechanically combined. Similarly, since crowd raters do not have impressions of a student's past performance history like faculty members might, there should be few concerns about bias based on previous evaluations. Crowdsourced ratings can also address challenges with faculty rater motivation. For example, faculty may be tempted to inflate ratings and give a student the "benefit of the doubt" because they must interact with him or her in the future. Since lay raters and students share no social connection, this concern is largely eliminated.

Fourth, when definitions of competent performance in a domain are divergent or dynamic, as they are in assessments of professionalism or communication skills, rater disagreement can be a meaningful and valuable form of feedback. Crowdsourcing may offer a way for students to receive both a larger volume of feedback as well as a more diverse, representative collection of perspectives regarding the efficacy of their behaviors. As the Mazor et al. (2007) study demonstrates, for every student behavior, there is a wide distribution of rater reactions. Different raters focus on and value different aspects of performance and disagree consistently about the appropriateness of the same behaviors. Similarly, practicing physicians see

a diverse set of patients that may have diverging opinions about effective and ineffective communication practices. Thus, medical students should be prepared early in their education to understand diversity in patient reactions and should have enough exposure to patient reactions to build a robust mental model of a reasonable range of physician behavior-patient response relationships. US-based crowd raters who see physicians are, by definition, part of the patient population American medical students might encounter as they enter practice (although it is possible the crowd is not necessarily a representative sample of the patient population). In light of Mazor et al.'s (2005) findings that OSCE communication ratings by trained-raters do not predict patient satisfaction scores, perhaps there is no more appropriate and relevant judge of communication and interpersonal skills than the ultimate consumer of those physician behaviors, the patient. In this sense, one could argue that a scoring solution based on patient proxy consensus is a more appropriate model than, for example, a true-score or gold-standard model driven by expert assessors.

Lay Rater Accessibility to Various Rating Tasks and Formats

Although Landy and Farr proposed a moratorium on rating format research in 1980 (Landy & Farr, 1980), citing evidence that the rating format only accounted for 4-8% of the variance in ratings, research on rating formats continued. Indeed, the choice of assessment instrument has significant implications for measuring complex clinical competence skills, such as communication. A checklist item such as "Be non-judgmental" is not a discrete behavior (e.g., "Washed hands") suited for a dichotomous performed/omitted decision on a checklist. It is instead a complex behavior that requires nuanced judgment across multiple dimensions (e.g., tone of voice, content of statements and questions). Although checklists may provide an "illusion of objectivity," they are notoriously unreliable and promote "shotgun behavior" rather than the

demonstration of competence and expertise (Norman, 2005). Furthermore, the selection of items into a checklist is an inherently subjective exercise dependent on judgments made by subject matter experts (Ilgen et al., 2015), which further undermines the notion that checklists are entirely objective assessments.

Some may question the necessity for multiple ratings of checklists that have discrete, "objective" items that are marked as either performed, not performed, or performed incorrectly because these actions are readily observable and thus should not be subject to much rater disagreement. Ilgen et al.'s review indeed found that the reliability coefficients for checklists were higher than in previous studies. One explanation for this is that many of the studies included in the Ilgen et al. review featured measurement of procedural skills. As such, the authors proposed that technical or procedural skills might lend themselves to more reliable measurement than broader competencies such as communication (Ilgen et al., 2015). On the other hand, Dicker et al. found that single trained-rater reliability coefficients on a dichotomously scored (performed/not performed) communication skills checklist were below acceptable levels for both SPs (.74) and non-SPs (.40).

Psychometric experts in the medical education community have advocated replacing skills checklists with global rating scales, which require the use of rater judgment (Norman, 2005; Regehr et al., 1998). The relative merits and limitations of checklists versus global rating scales have long been a source of scholarly debate, but a recent meta-analysis by Ilgen et al. (2015) quantitatively summarized and compared the reliability of both types of instruments in simulation-based assessments. Although checklists and global rating scales shared on average 58% of the variance, "compared with the checklist, the GRS has higher average inter-item and inter-station reliability, can be used across multiple tasks, and may better capture nuanced

elements of expertise." While global rating scales may be more useful than checklists once students have developed some domain-relevant expertise, checklists provide a clear, unified set of criteria and may be more appropriate for assessing and providing feedback to early learners on sequenced tasks such as physical exams and procedures.

Despite having recorded students performing physical examinations as part of the OSCE, Mazor et al. (2007), excluded lay participants from rating the physical examination checklist component because, "lay raters were not considered sufficiently knowledgeable to judge the appropriateness or technique of the [physical] examination." However, lay raters were able to converge with expert ratings on arguably more technical tasks in the surgical crowdsourcing studies. The pivotal difference here then may lie not in the degree of sophistication of the target behavior, but rather in our expectations of raters and the rating format. Specifically, the crowdsourcing studies asked raters to make global judgments that may be more accessible to lay raters when compared with detailed physical exam checklists, as in the Mazor et al. (2007) study, which are typically loaded with technical language that lay raters might have trouble understanding without training (e.g., "Auscultate for heart sounds in the aortic, pulmonic, tricuspid and mitral areas of the heart using the bell of the stethoscope"). To my knowledge, no study of crowdsourcing or lay raters has compared lay rater performance on technical checklists versus a global rating scale.

Student Reactions to Crowdsourced OSCE Ratings

Although several studies described above have investigated medical student reactions to OSCE evaluations by student peer examiners, no study to my knowledge, including the surgical crowdsourcing work, has measured trainee reactions to assessment ratings provided by lay raters. Student buy-in to the appraisal system is a must if educators are to use crowdsourced ratings and

feedback to motivate student behavioral change. For example, students must have trust towards the evaluator. They must believe that crowd raters, as a group, are adequate judges of their standing in a particular domain (e.g., cognitive trust) and that crowd raters are unbiased and not motivated by ill intent (e.g., affective trust). Students must perceive crowd feedback as timely, specific, high quality, and useful for fostering improvement (i.e., utility). Students may also view fairness and utility differently depending on the use of the crowdsourced assessment data. For example, students may support such a system if it were used for formative purposes but may be reluctant to trust crowdsourced ratings in summative applications.

Additionally, students may have concerns about privacy issues if their performance videos are exposed to unknown lay raters around the world. Although laparoscopic surgery videos are naturally de-identified because rating tasks typically present raters with an intracorporeal camera view, OSCE performances require additional processing to blur faces in order to protect student identity. Recent advances in video processing software (e.g., YouTube's Blur All Faces; Filmora's Face Off) have facilitated rapid and automatic face blurring. Unfortunately, blurring faces constricts the social information channel and thus reduces the amount of behavioral information available for raters to make judgments about communication and interpersonal skills; however, protecting student privacy is not only a priority but a legal requirement per FERPA (the Family Educational Rights and Privacy Act).

Research Questions

With the above considerations in mind, this study aims to address the following research questions (RQ):

- RQ 1: How accurate and reliable are crowd ratings when compared with SP and faculty ratings in evaluations of clinical competencies elicited in OSCE-style assessments?

- RQ 2: How many crowd ratings are required to reach reliability levels that match or exceed SP and faculty ratings?
- RQ 3: What is the difference in latency in collecting ratings from SPs, faculty, and crowd workers?
- RQ 4: How much does it cost to collect an accurate and reliable set of crowdsourced ratings?
- RQ 5: Does the rating format (e.g., technical checklist vs. global rating scale) make a difference with respect to the accuracy and reliability of crowdsourced ratings?
- RQ 6: How do medical students react to crowdsourced OSCE ratings across a variety of criteria, including privacy, quality, utility, and trust?

In order to ease interpretation, this investigation was divided into two parts. Study 1 addressed RQ 1 – 5, and Study 2 dealt with RQ 6. All study procedures were approved by the University of South Florida and Rush University Medical Center Institutional Review Boards (Appendix 1).

CHAPTER THREE:

STUDY 1 METHOD

The name of each rater type is capitalized when referring to raters who participated in this study (e.g., Crowd raters, SP raters, Faculty raters). When referring to these types of individuals in general, I follow typical capitalization conventions.

Participants

Standardized patient raters. Four actors, who regularly serve as standardized patients at Rush Medical College (RMC), volunteered to serve as SP raters. These actors were different actors from the ones used to *portray* the standardized students and the standardized patients in the video recordings of the clinical scenarios.

All SP raters had received training in using the RMC Interpersonal and Communication Skills Form (ICSF, Appendix 2) and the RMC Physical Exam Skills Checklist (PESC, Appendix 3). SPs at RMC are trained to use the ICSF and PESC tools through PDT and FOR procedures. The average experience of the SP raters at any institution ranged from 2 to 15 years (M = 8.75, SD = 6.24) and, at RMC specifically, from 2 to 4 years (M = 3.25, SD = .96). All SP raters estimated they had participated in approximately 100 one-on-one student encounters at any institution and between 30 to 100 encounters at RMC (M = 48.75, SD = 34.25). Of all encounters at any institution, SP raters estimated that 75% to 90% (M = 82.50%, SD = 8.66%) included evaluating interpersonal, social, or communication skills, 50% to 75% (M = 63.75%, SD = 11.09%) included evaluating physical examination skills, and 15% to 50% (M = 31.25%, SD =

18.87%) included evaluating cardiovascular exam skills, specifically. Of all encounters at RMC, SP raters estimated that 75% to 80% (*M* = 76.25%, *SD* = 8.66%) included evaluating interpersonal, social, or communication skills, 25% to 75% (*M* = 55.00%, *SD* = 22.73%) included evaluating physical examination skills, and 15% to 25% (*M* = 18.75%, *SD* = 4.79%) included evaluating cardiovascular exam skills, specifically. One of the four SP raters had served as an SP trainer, who taught other SPs to evaluate interpersonal, social, communication, physical exam, and cardiovascular exam skills.

Faculty raters (ICS task). Four faculty members from RMC's behavioral science faculty (BHVFac) volunteered to serve as Faculty raters for the Interpersonal and Communication Skills (ICS) task. These faculty members typically hold doctoral degrees in clinical psychology, have expertise in clinical communication techniques and rapport development, teach patient interviewing curriculum at RMC, and regularly serve as Faculty raters for the ICS component of RMC OSCEs. BHVFac raters at RMC are trained to use the ICSF through PDT and FOR procedures. The average experience of the BHVFac raters teaching in the RMC Patient Interviewing course, which is where RMC students learn appropriate patient-facing social and communication skills, ranged from 2 to 12 years (*M* = 7.25, *SD* = 4.11). BHVFac raters reported evaluating between 110 and 1028 (*M* = 609.00, *SD* = 379.90) student videos for interpersonal, social, and communication skills as part of their work in the RMC Patient Interviewing Course.

Faculty raters (PE task). Four physician faculty members (PhysFac) volunteered to rate the Physical Exam (PE) task. Although these faculty members do not typically evaluate students in OSCEs, they can be considered socially nominated "experts" in basic physical exam procedures. PhysFac raters had practiced medicine between 5 and 18 years (*M* = 11.00, *SD* =

5.35). PhysFac raters reported teaching between 0 and 425 students, residents, or other healthcare professionals to perform a cardiovascular exam (M = 116.75, SD = 206.32) and evaluated between 0 and 128 students, residents, or other healthcare professionals performing a cardiovascular exam (M = 38.50, SD = 60.25). One of the PhysFac raters, the one with the most experience teaching and evaluating cardiovascular exam, had one year of teaching experience in the RMC Physical Diagnosis course, which is where students first learn physical exam skills.

Crowd raters. Crowd raters were recruited from Amazon's Mechanical Turk (MTurk) crowdsourcing utility. Crowd raters were screened into the study if they: 1) had completed at least 500 Human Intelligence Tasks (the unit of work in MTurk) with a 97% approval rate, and 2) resided in the US and had seen a US-based doctor as a patient or as the guardian of a patient within the last two years. The first set of screening criteria were used to control the quality of MTurk workers (Sheehan & Pittman, 2016). The second set of criteria was used to ensure that Crowd raters had an understanding of the norms concerning the behaviors and practices of US-based physicians.

In total, 150 unique Crowd raters completed 239 rating tasks. 27% of Crowd raters completed more than one rating task. Of those raters who completed more than one rating task, 44% completed 2 ratings, 17% completed 3 ratings, 20% completed 4 ratings, 17% completed 5 ratings, and 2% completed 6 ratings.

The age of Crowd raters ranged from 20 to 70 years (M = 36.75, SD = 11.11). Crowd raters were 57% female and 43% male. With respect to race, 77% of Crowd raters were White, 9% Black/African American, 6% Hispanic/Latino, 4% Asian, 1% American Indian/Alaska Native, and 2% Other. 33% of Crowd raters achieved a Bachelors degree as their highest level of education, 23% attended some college but did not receive a degree, 14% earned an Associate

degree, 13% graduated high school, 10% held an advanced degree (Masters, PhD, or MD), 4% had finished trade/technical school, and 2% had not earned a high school degree. Of those 12 Crowd raters who reported having an occupation related to healthcare, 6 had occupations in the healthcare industry that involved the direct provision of patient care.

Crowd raters were compensated 15 cents per estimated task minute, which is consistent with the ethics and quality recommendations from Sheehan & Pitmman (2016). This wage comes out to $9/per hour, which is $1.75/hour above the current U.S. federal minimum wage, $0.75/hour above the Illinois State minimum wage, and $1.50/hour below the City of Chicago minimum wage.

Measures

Two instruments were used to rate OSCE performance, both of which were developed in-house at RMC. The Interpersonal and Communication Skills Form (ICSF; Appendix 2) is designed to measure patient-centered communication behaviors involved in patient interviewing, consultation, and interactions with families. The domains were derived from those outlined by United States Medical Licensing Exam (USMLE) Content Outline ("USMLE content outline," n.d.). Although these skills are assessed by the USMLE Step 2 Clinical Skills licensing exam, the National Board of Medical Examiners, the agency which administers this OSCE-style assessment, does not provide the items or rubrics their raters use to judge performance. Therefore, schools have had to rely on the content outline to construct their own instruments to evaluate these behaviors. RMC's ICSF is a mixed-form instrument that uses a combination of six yes-no checklist items that assess how students begin and end patient encounters, eight five-point behaviorally-anchored global items that assess more complex behaviors (e.g., Using Appropriate and Sensitive Language, Empathy for the Patient's Distress, Supporting Emotions & Fostering

Relationship), and two open-ended items that are intended to provide students with narrative feedback about what they did well in the simulated patient encounter and what they could improve in future encounters. Lower scores on the ICSF are indicative of poorer or more maladaptive communication behaviors and higher scores are indicative of more adaptive, desirable communication skills.

The Physical Exam Skills Checklist (PESC; Appendix 3) measures student performance on a set of physical examination maneuvers. The full version of the PESC captures all possible maneuvers in a head-to-toe physical exam. However, rather than performing full head-to-toe examinations, physicians will typically perform specific sets of physical exam maneuvers in their exploration of a diagnostic hypothesis informed by the patient's chief complaint, history of present illness, a review of systems, and other pertinent information. As such, the PESC is modularized so that evaluators can pull out subsets of items, typically grouped by organ systems (e.g., cardiovascular, neurological, respiratory). The PESC typically uses a three-point scale, "Done", "Done Incorrectly", "Not Done" with a written explanation required for each maneuver marked as "Done Incorrectly". In application, the PESC is filled out by an SP immediately after a live encounter with a student. However, since this study used video-recorded OSCE performances, the PESC was adapted to include a fourth option, "Observation Obscured," to account for those times when video production obscured a critical portion of the maneuver such that it was difficult for the rater to be sure whether the student performed the maneuver correctly or not. Selecting the "Observation Obscured" option also required raters to provide a written justification.

Two versions of the Rater Reactions Questionnaire, one for the ICS task (RRX-ICS; Appendix 4) and another for the PE task (RRX-PESC; Appendix 5), were used to measure crowd

worker perceptions of the evaluation tasks. The RRX was based on Moineau et al. (2010) and featured items such as "I felt comfortable evaluating this student's physical exam skills" and "I feel like I need additional training to make such evaluations in the future." The questionnaires were identical, except for the skill type each item referred to (i.e., physical exam skills vs. interpersonal and communication skills) and two questions about narrative feedback were removed from the RRX-PESC since raters were not asked to leave narrative feedback on the PESC. The questionnaire used a 6-point Likert-type scale, which asked raters to select how much they agreed with each item (1 = Strongly Disagree, 6 = Strongly Agree).

Demographic questionnaires for all rater types are presented in Appendix 6.

Video Stimuli

Six videos were recorded and used as OSCE performance episodes for raters to evaluate. Because it was unclear how students might react to a crowdsourced OSCE rating system, particularly with respect to privacy issues, this study used "standardized students," portrayed by actors typically employed to portray standardized patients. Six actors (4 female, 2 male) played the roles of standardized students. Two additional actors (1 female, 1 male) played the role of standardized patients (3 videos each). The videos were recorded in a traditional patient exam room using 2-3 camera angles typical of OSCE events. After recording, the video stimuli were separated into two components, an interview and plan clip designed for the ICS task and a physical exam clip designed for the PE task; this process yielded a total of 12 videos. The videos were edited to present all available camera angles simultaneously in order to maximize the rater's visibility of the patient exam room. All videos were processed to blur the standardized student's face in order to simulate the privacy measures that would likely be applied in a crowdsourced OSCE rating system using actual medical students. Finally, the videos were

uploaded to a private YouTube channel so they could be embedded in the rating tasks. See Appendix 7 for a sample frame from each video type.

The clinical scenario for the videos was selected by the RMC Director of Clinical Communication Training and Research. The case featured a chief complaint commonly encountered by medical students (chest pain) and elicited the performance of typical OSCE station tasks – reviewing the patient's chart (see sample door chart in Appendix 8), collecting a history of present illness and focused social history, performing a focused cardiovascular physical exam, providing the patient with a preliminary assessment, and negotiating a preliminary treatment plan with a patient.

Each of the standardized student actors based their performance on scripts intended to model behaviors across the performance continuum. The case blueprint, the ratings tools (ICSF, PESC), and the judgment and experience of the RMC Director of Clinical Communication Training and Research were leveraged to generate two scripts each of behaviors typical of low-, intermediate-, and high-skilled students for a total of six scripts. Although the scripts were generated using a top-down approach to establish levels of performance a priori, there was natural variation in actor script interpretation and performance delivery. Therefore, a bottom-up approach was also used in post-production to benchmark the skill level in each video. The RMC Manager of Simulation Education, the final adjudicator in the RMC curriculum of rating discrepancies for OSCE events, reviewed all of edited videos and provided ratings using the ICSF and the PESC. These ratings are considered the gold standard, or true score, ratings in this study.

The interview and plan video segments ranged from 5 minutes and 20 seconds to 12 minutes and 51 seconds. The length of these videos tended to co-vary with the ground truth

ratings such that longer videos were associated with better performance on the ICSF. The physical exam video segments ranged in length from 1 minute and 42 seconds to 3 minutes. Table 1 provides a summary of the properties of the video stimuli and descriptive statistics for the associated rating tasks.

Collection of Ratings

All ratings tasks were developed and deployed through Survey Gizmo, a Web-based survey tool. The first page of each rating task featured the screening questions appropriate to each rater type, the second page featured the consent form, and instructions for the rating task on the third page. All video segments were embedded in the rating task, followed immediately by the appropriate tool (ICSF, PESC). All raters were provided with instructions on the technical aspects of video review (e.g., full screen mode, volume adjustment, reviewing segments), asked not to distribute the video links, and instructed to watch in full each video for which they would be providing ratings. A schematic of the ratings collected from each rater source is presented in Table 2.

Faculty and SP ratings. All faculty and SP raters were invited to participate in the rating task via email. Although it would have been ideal to incorporate the faculty and SP rating tasks into the normal curricular evaluation cycle in order to approximate the motivation conditions raters experience when reviewing real student performance, this was not possible due to logistical constraints. Specifically, SP raters always provide ratings immediately after a live encounter (not using video review) and might be familiar with the voices of their actor colleagues portraying standardized students, PhysFac raters do not typically serve as raters of OSCE encounters in the RMC curriculum, and BhvFac raters have a predetermined student case

load, know which cases are part of the present evaluation (the present chest pain case was not one of them), and also might be familiar with the voices of the standardized student actors.

PhysFac raters viewed all six physical exam video segments and provided ratings for each using the PESC; a task overview for PhysFac raters is provided in Appendix 9. BhvFac raters viewed all six interview and plan video segments and provided ratings for each using the ICSF; a task overview for BhvFac raters is provided in Appendix 10. SP raters viewed all twelve video segments and provided ratings for each using the PESC and ICSF; a task overview for SP raters is provided in Appendix 11. All video segments within rater were randomized in order to control for order effects. After providing ratings, faculty and SP raters filled out the demographics questionnaire corresponding to their rater type.

Crowd ratings. Two separate rating tasks were developed for the Crowd raters, one for reviewing the interview and plan video segments using the ICSF and a second for reviewing the physical exam video segments using the PESC. A PE task overview for Crowd raters is provided in Appendix 12, and an ICS task overview is provided in Appendix 13. For the ICS task, Crowd raters were provided a set of written instructions beyond those provided to the SP and Faculty raters. Notably, Crowd raters were provided with 1) background about who standardized patients are and the role of standardized patient encounters in providing feedback to medical students, 2) instructions about how to interpret and apply the behavioral anchors for the global ICSF items, 3) a prompt about the importance of being accurate and honest in their ratings and feedback, and 4) a prompt to put themselves in the standardized patient's shoes as they complete the ICSF. The Crowd rater screening page used a Captcha verification to ensure that the task was being completed by a human, not a machine. After providing ratings using the ICSF, Crowd raters completed the RRX-ICSF and Crowd rater demographics form.

Multiple attention checks were embedded in the ICS task. A six-item true/false attention check quiz verified that Crowd raters read and understood the instructions. Next, three items verified that the Crowd rater reviewed the video: 1) a multiple-choice item about the patient's chief complaint, 2) an open-ended short answer item about the factors that aggravate the patient's condition, and 3) an open-ended short answer item about the medical student's and attending physician's recommended treatment plan. Finally, one item embedded in each of the instruments (ICSF, RRX-ICSF) ensured that Crowd raters were not haphazardly responding (e.g., "If you're not a robot, select Strongly Disagree for this item").

For the PE task, Crowd raters were provided a set of written instructions beyond those provided to the SP and Faculty raters. Notably, Crowd raters were provided with 1) background about who standardized patients are and the role of standardized patient encounters in providing feedback to medical students, 2) a 3-minute training video walking the rater through the proper technique for a cardiovascular exam, 3) instructions that all exam maneuvers are to be performed on skin and that sports bras are considered skin for female standardized patients, and 4) a series of visual guides clarifying elements specific to a cardiovascular exam (e.g., the location of the ulnar surface of the hand, the location of each of the heart areas, the difference between the bell and diaphragm of a stethoscope). The Crowd rater screening page used a Captcha verification to ensure that the task was being completed by a human, not a machine. After providing ratings using the PESC, Crowd raters completed the RRX-PESC and the Crowd rater demographics form.

Multiple attention checks were embedded in the PE task. A ten-item true/false attention check quiz verified that Crowd raters read and understood the instructions. Next, one multiple choice item asked which type of physical exam was being evaluated in the encounter. Finally,

one item embedded in each of the instruments (PESC, RRX-PESC) ensured that Crowd raters were not haphazardly responding (e.g., "Select 'Visible Not Done' for this item").

Twenty assignments for each video review task were posted as Human Intelligence Tasks (HITs) on Amazon's Mturk crowdsourcing engine (see Appendix 14 for a sample HIT). In total, 240 assignments (12 videos x 20 rating assignments per video) were posted. Although raters could only complete one assignment within a HIT (i.e., they could only provide ratings and feedback for the same video once), raters were allowed to sign up for as many different videos as they wanted. This configuration allowed for the approximation of conditions if a crowdsourced rating system were to be used at scale with real students. Permitting Crowd raters to review multiple videos allows for the natural development of Crowd rater expertise as well as an increased video review rate since the rater pool is left unrestricted.

Based on pilot testing, the ICS task was estimated to take approximately 20 minutes plus the time of video review to complete. Considering the length of each video varied, the compensation for completing each task varied as well. Since the differences in video segment time for the PE task were nominal, each PE task was estimated to take approximately 15 minutes to complete. Crowd raters were compensated based on the estimated total task time. Amazon charges a fee to use its MTurk crowdsourcing engine. For HITs with less than 10 assignments, Amazon charges a 20% fee on top of the compensation paid to the rater; for HITS with 10 or more assignments, Amazon charges a 40% fee. Since all video review HITs in this study provided 20 assignments, Amazon charged a 40% fee.

CHAPTER FOUR:

STUDY 1 RESULTS

Data Preparation

Although 240 total rating assignments were posted, one MTurk worker accepted a rating assignment but never participated in the study. This assignment was rejected and, due to a technical error on MTurk, was never re-posted. This brought the total amount of ratings to 239.

All attention checks items were inspected to identify haphazard responding. If any ratings met at least one of the following criteria, they were excluded from analyses. The number of ratings excluded based on each condition is indicated in parentheses. For the ICS task: 1) incorrect identification of the chief complaint (n = 0), 2) incorrect identification of aggravating conditions (n = 2), 3) incorrect identification of the treatment plan (n = 1), 4) incorrect response to the attention check items embedded in the ICSF or RRX tools (n = 2), or 6) incorrect response to more than one of six instruction questions (n = 3). One rating qualified for exclusion based on two criteria, so 7 ratings were excluded for the ICSF task. For the PE task: 1) incorrect identification of the physical exam type (n = 2), 2) incorrect response to the attention check items embedded in the PESC or RRX tools (n = 9), and 3) incorrect response to more than two of ten instruction questions (n = 2). One rating qualified for exclusion based on two criteria, so 12 ratings were excluded for the PE task.

Within a typical MTurk HIT, it is the right of the requester to reject work not completed to standards. However, research ethics dictate that participants be compensated irrespective of

their responding patterns. As such, MTurk raters who were excluded were still compensated for their participation.

The surgical crowdsourcing studies (e.g., Chen et al., 2015; Holst et al., 2015), treated rating results at the scale level (e.g., composite scores of multiple global ratings). While this strategy may be appropriate when assessments are used to make promotion decisions wherein good performance on one item may compensate for poor performance on another, ignoring item-level information when the assessment is being used for developmental feedback is inappropriate. It is at the analytic, rather than global or composite, level where educators can provide feedback that trainees can use to drive specific behavioral change. As such, I use analyses that prioritize item-level information.

Crowd Demographics

Provided that Crowd raters acted as patient proxies in this study, it is important to consider how well the Crowd raters represent the demographics of the US patient population. Table 3 compares the demographic characteristics (gender, age, education marital status, race, income) of the Crowd raters in this study with the demographic characteristics of the US population as estimated by the 2015 1-Year Current Population Survey (US Census BureauUS Bureau of Labor Statistics, n.d.). Pediatric patients are excluded from analysis because, per MTurk's terms of use, only adult participants took part in this study. In comparison with the US population, more Crowd raters were female and single. The Crowd raters were also younger, achieved higher levels of education, and had lower combined household incomes. With respect to race, Crowd raters were about as diverse as the US population with both groups predominantly White (Crowd = 77%, US = 76%).

Timing and Cost

All timing data presented below is accompanied by projected costs for each rater type per rating. According to the US Bureau of Labor Statistics, clinical psychologists employed in a hospital setting earned a median salary of $91,180 (US Bureau of Labor Statistics, n.d.). Benefits account for approximately 30% of total compensation in private industry (US Bureau of Labor Statistics, 2016), which brings total compensation of a clinical psychologist to $130,257, an hourly rate of $62.62 ($1.04/minute). Physicians and surgeons earned a median salary of $187,200 (US Bureau of Labor Statistics, n.d.), which equates to a total compensation of $267,428, and an hourly rate of $128.57 ($2.14/minute). SPs at RMC are compensated as contractors (i.e., no benefits) at $22/hour ($0.37/minute). Crowd raters were compensated in this study at $9/hour ($0.15/minute) plus a 40% MTurk fee of $3.60/hour ($0.06/minute) when collecting more than 9 ratings per video. This brings the total Crowd rater rate to $12.60/hour ($0.21/minute). For the reliability analyses presented below where fewer than 9 Crowd raters are included, this lowers the MTurk fee to 20% ($1.80/hour or $0.03/minute) for a total Crowd rater fee of $10.80/hour ($0.18/minute).

The average time it took each rater type to complete each type of rating task by video is presented in Table 4. For Crowd raters, only the first time they participated in each task type was retained in analyses to approximate the time it would take a new rater to complete the task. After all of the within-task repeats were excluded, time outliers (analyzed within-task, within-rater type, within-video) were removed through a visual examination of box plots.

Two task-time metrics are presented. The first metric accounts for the time it took a rater to complete the entire rating task (Total Task Time). This was calculated from the total time spent on the survey page containing the video and the ICSF or PESC. We can be sure that this

number represents task time only (and not screening time, consent time, etc), because participants were only allowed to forward navigate in the survey. There were occasional extreme values for Total Task Time (e.g., 1.5 hours), but these were most likely due to raters leaving the survey open on their device while they engaged in other activities. These extreme values were removed through the outlier analysis. The survey page differed between Crowd raters and SP or Faculty raters since Crowd raters had instructions included on their page, whereas SP and Faculty raters had instructions included on a separate page. This approximates conditions in a realistic crowdsourced OSCE rating solution if operating under the assumption that a rating task might be picked up by a Crowd rater engaging in this type of rating activity for the first time. With respect to mean Total Task Time for ICS (all times are presented as mm:ss), BhvFac raters were the fastest (*M* = 10:20; $10.74), followed by the SP raters (*M* = 17:16; $6.39), and the Crowd raters (*M* = 21:30, $4.52). For the PE task, SP raters had the shortest mean Total Task Time (*M* = 5:22; $1.99), followed by the PhysFac raters (*M* = 6:04; $12.99), and the Crowd raters (*M* = 15:26; $3.24).

The second task-time metric holds video review time constant (Instrument Time). This metric is calculated by subtracting the video length from the Total Task Time, which theoretically yields the time it took each rater to complete the instrument if he or she *at least* watched the video all the way through a single time. With respect to mean Instrument Time for ICS, BhvFac raters were the fastest (*M* = 2:05; $2.16), followed by the SP raters (*M* = 9:18; $3.44), and the Crowd raters (*M* = 13:15; $2.78). For the PE task, SP raters had the shortest mean Instrument Time (*M* = 2:44; $1.01), followed by the PhysFac raters (*M* = 3:26; $7.34), and the Crowd raters (*M* = 12:48; $2.69).

From a theoretical perspective, it may be reasonable to assume that the video review time might vary across rater types despite the same video length. For example, Crowd raters may spend more time than SP or Faculty raters re-reviewing video since they may not have a robust mental model of the association between student behaviors and items on the PESC or ICSF. Since video review time in this study is confounded with instrument time in the Total Task Time metric, it is not possible to determine how long raters spent actually viewing each video as opposed to filling out the instrument. However, one piece of data suggests that some raters may not be reviewing the entire video. Specifically, the mean Instrument Time for BhvFac raters for the DBRL ICS video was 4 seconds (mean Total Task Time = 12:55; video length = 12:51) and the Instrument Time for PhysFac raters for the BKRL PE video was 9 seconds (mean Total Task Time = 3:25, video length = 3:16). This suggests that Faculty raters may not have watched the video in its entirety, that they rushed through the instrument, or perhaps that they fill out the instrument and watch the video simultaneously. This may be the case for other rater types as well, but it is not possible to tell since video review time is confounded with instrument time in the Total Task Time metric.

The total time to collect each package of ratings is presented in Table 5. Although the ICS and PE tasks were deployed at separate times on MTurk, all assignments within a task were deployed simultaneously. With all ICS tasks deployed at the same time, the minimum amount of time it took to collect 20 ICS ratings for a single video was 2 hours and 58 minutes, while the longest amount of time was 5 hours and 43 minutes. With all PE tasks deployed at the same time, the minimum amount of time it took to collect 20 PE ratings for a single video was 1 hour and 32 minutes, while the longest amount of time was 3 hours and 33 minutes. It took 16 days, 12 hours, and 29 minutes to collect the entire package of ICS ratings (1 set of ratings for 6 videos from 4

BhvFac raters). It took 8 days, 18 hours, and 56 minutes to collect the entire package of ICS ratings (1 set of ratings for 6 videos from 4 PhysFac raters). It took 10 days, 4 hours, and 52 minutes to collect the entire package of ICS and PE ratings (1 set of ratings for 6 videos from 4 SP raters).

Rating Distribution and Mean Performance Levels

The distribution of responses (item x video x rater type) on the ICSF is presented in Table 6 for checklist items and in Table 7 for global (Likert-type) items. Table 8 presents mean ratings for all ICSF items (item x video x rater type). Figure 1 presents mean ratings as videos within rater type while Figure 2 presents mean ratings as rater type within video for global items. Figure 3 presents mean ratings as videos within rater type for global items while Figure 4 presents mean ratings as rater type within video for checklist items.

A one-way ANOVA was conducted to determine if mean ICSF checklist ratings differed by rater type. Raters were classified into three groups by rater type: Crowd (n = 678), Faculty (n = 144), and SP (n = 144). ICSF checklist ratings were statistically significantly different by rater type, $F(2, 963) = 10.87$, $p < .0005$, partial $\eta^2 = .02$. Crowd raters provided the highest checklist scores (M = .86, SD = .35), followed by SP raters (M = .78, SD = .42), and Faculty raters (M = .70, SD = .46). Tukey post hoc analysis revealed that the mean difference between Crowd and Faculty raters (.15, 95% CI [.07, .24] was statistically significant ($p < .0005$). No other group differences were statistically significant.

A one-way ANOVA was conducted to determine if mean ICSF global ratings differed by rater type. Raters were classified into three groups by rater type: Crowd (n = 904), Faculty (n = 192), and SP (n = 192). ICSF global ratings were statistically significantly different by rater type, $F(2, 1285) = 71.97$, $p < .0005$, partial $\eta^2 = .10$. Crowd raters provided the highest global ratings

(*M* = 3.73, *SD* = 1.20), followed by SP raters (*M* = 3.34, *SD* = 1.05), and Faculty raters (*M* = 2.63, *SD* = 1.22). Tukey post hoc analysis revealed that the mean difference between Crowd and Faculty raters 1.11, 95% CI [.89, 1.33] was statistically significant ($p < .0005$). The mean difference between Crowd and SP raters .39, 95% CI [.17, .61] was statistically significant ($p < .0005$). The mean difference between SP and Faculty raters .71, 95% CI [.43, 1.00] was statistically significant ($p < .0005$).

Because the PESC uses a nominal scale, only descriptive statistics in the form of a distribution of PESC ratings (item x video x rater type) is presented in Table 9.

Accuracy

Rater accuracy for the ICS task was analyzed using two indices: proportion of raw agreement (RA) and Average Deviation (AD; Burke, Finkelstein, & Dusig, 2016). These indices were selected due to their common-sense value and high interpretability. The ratings provided by the Manager of Simulation Education Manager served as the true score. RA was defined as the proportion of cases for which the rater agreed with the true score rating exactly. AD, calculated only for the ICSF global items, was defined as the average of the rater's rating minus the true score rating. Two variants of AD are presented, one which captures only absolute deviation (ADabs) and a second that captures the vector or direction of deviation (ADvec).

ICSF RA values (rater x item) are presented in Table 10. For the six ICSF checklist items, Faculty and SP raters had the highest RA values for one item, Crowd raters had the highest for two, and SPs had the highest for three items. At the item level, SP and Faculty raters were tied for the highest RA for one of the eight global ICSF items. For the remaining seven items, Crowd raters had the highest RA for one item and SP raters and Faculty raters had the

highest RA values for three items each. Mean RA (across raters) for checklist items ranged from .53 to .99 while mean RA for global items ranged from .28 to .51.

Averaging across items, differences in the agreement proportions across rater type were explored using the chi-square test of homogeneity. With respect to RA for the checklist items of the ICSF, Crowd ratings agreed with true score ratings 83.2% of the time, Faculty ratings agreed 74.3% of the time, and SP ratings agreed 86.1% of the time, a statistically significant difference in proportions, $p = .02$. Post hoc analysis involved pairwise comparisons using the z-test of two proportions with a Bonferroni correction. The proportion of agreement for Faculty raters was statistically significantly lower than for the SP and Crowd raters, $p < .05$. The proportions of agreement for the SP and Crowd raters was not statistically significantly different, $p > .05$.

When grouped by the Skill Level of the student in the video, Crowd ratings for High Skill videos for ICS checklist items agreed with true score ratings 90.2% of the time, Faculty ratings agreed 83.3% of the time, and SP ratings agreed 95.8% of the time, which was not a statistically significant difference in proportions, $p = .12$. Crowd ratings for Low Skill videos for ICS checklist items agreed with true score ratings 79.5% of the time, Faculty ratings agreed 69.8% of the time, and SP ratings agreed 81.3% of the time, which was not a statistically significant difference in proportions, $p = .08$.

With respect to RA for the global items on the ICSF, Crowd ratings agreed with true score ratings 28.7% of the time, Faculty ratings agreed 44.8% of the time, and SP ratings agreed 44.3% of the time, a statistically significant difference in proportions, $p < .0005$. Post hoc analysis involved pairwise comparisons using the z-test of two proportions with a Bonferroni correction. The difference in proportions of agreement for Crowd raters was statistically

significantly lower than for the SP and Faculty raters, $p < .05$. The difference in proportions of agreement for the SP and Faculty raters was not statistically significantly different, $p > .05$.

When grouped by the Skill Level of the student in the video, Crowd ratings for High Skill videos for ICSF global items agreed with true score ratings 38.5% of the time, Faculty ratings agreed 37.5% of the time, and SP ratings agreed 45.3% of the time, which was not a statistically significant difference in proportions, $p = .56$. Crowd ratings for Low Skill videos for ICS global items agreed with true score ratings 28.7% of the time, Faculty ratings agreed 44.8% of the time, and SP ratings agreed 44.3% of the time, a statistically significant difference in proportions, $p < .0005$. Post hoc analysis involved pairwise comparisons using the z-test of two proportions with a Bonferroni correction. The proportion of agreement for Crowd raters was statistically significantly lower than for the SP and Faculty raters, $p < .05$. The proportions of agreement for the SP and Faculty raters was not statistically significantly different, $p > .05$.

One additional chi-square test explored rating integrity for all rater types. Specifically, one set of three checklist items appears at the beginning of the ICSF tool (Beginning the Encounter) and three additional checklist items appear at the end of the ICSF (Ending the Encounter). The percentage of raw agreement with true scores for the first set of checklist items was 91.7% while the raw agreement for the set of items at the end of the tool was 72.9% across all raters, a statistically significant difference, $p < .0005$. This suggests that raters agreed more with true scores for behaviors and checklist items at the beginning of the video and instrument when compared with those at the end. This is perhaps a function of 1) rater fatigue, 2) the items at the end of the encounter being more difficult to evaluate, or 3) raters not watching each video in full.

ADabs and ADvec values (item x rater type) for ICSF global items are presented in Table 11. The general trend for ADabs was that Crowd raters (M = 1.06) deviated more from true scores than Faculty (M = .64) or SP raters (M = .69). With respect to ADvec, Crowd raters (M = .82) and SP raters (M = .46) were more lenient relative to true scores when compared with Faculty raters (M = -.25)

ADabs and ADvec analyses were split by Skill Level (High, Low) and Rater Type (SP, Faculty, Crowd). Table 12 presents mean levels of ADabs and ADvec with a one-sample test for each value testing the null hypothesis that the average deviation from the true score equals zero. All but one of these values (ADvec for SP ratings of High Skill videos) was significantly different from zero.

A two-way ANOVA was used to explore whether the vectored deviations from true score were significantly different across Skill Level within Rater Type. The interaction between Rater Type and Skill Level was statistically significant, $F(2, 1282) = 8.06$, $p < .0005$, partial $\eta^2 = .01$.

A simple main effects analysis of Skill Level revealed that for Crowd Raters, mean deviation for High Skill videos was .29 (SD = .99) vs. 1.09 (SD = .59) for Low Skill videos, a mean difference of .80, 95% CI [.67, .94], $F(1, 1282) = 134.13$, $p < .0005$, partial $\eta^2 = .10$, which was statistically significant. Thus, Crowd raters viewing High Skill videos exhibited lower levels of deviation from true score than Crowd raters viewing Low Skill videos.

For Faculty raters, mean deviation for High Skill videos was -.39 (SD = .95) and -.18 (SD = .88) for Low Skill videos, a mean difference of .21, 95% CI [-.09, .51], $F(1, 1282) = 1.94$, p = .16, partial $\eta^2 = .00$, which was not statistically significant. Thus, Faculty raters viewing High and Low Skill videos did not differ in the deviations of their ratings from true score.

For SP raters, mean deviation for High Skill videos was .20 (*SD* = .95) and .59 (*SD* = .81) for Low Skill videos, a mean difference of .39, 95% CI [.09, .69], $F(1, 1282) = 6.66$, $p = .01$, partial $\eta^2 = .01$, which was statistically significant. Thus, SP raters viewing High Skill videos exhibited lower levels of deviation from true score than SP raters viewing Low Skill videos.

Rater accuracy for the PE task was analyzed using the RA procedure described above. AD analyses for the PE were not conducted because the PESC uses a nominal response scale. Differences in the agreement proportions across rater types were explored using the chi-square test of homogeneity.

RA values (item x rater type) for the PESC are presented in Table 13. Across 16 PESC items, 1 item had equal levels of RA across raters, 3 items had Faculty and SP raters with the same highest levels of RA, 1 item had Crowd and SP raters with equal highest levels of RA, 1 item with Crowd raters highest, 2 items with SPs highest, and 8 items with Faculty highest. Individual items ranged in RA (across raters) from .40 to .99.

Aggregated across items, Crowd PESC ratings agreed with true score ratings 59.7% of the time, Faculty ratings agreed 68.8% of the time, and SP ratings agreed 65.4% of the time, a statistically significant difference in proportions, $p < .0005$. Post hoc analysis involved pairwise comparisons using the z-test of two proportions with a Bonferroni correction. The proportion of agreement for Faculty raters was statistically significantly higher than for Crowd raters, $p < .05$. The differences in proportions of agreement between SPs and Crowd raters and between SPs and Faculty raters was not statistically significantly different, $p > .05$.

When grouped by the Skill Level of the student in the video, Crowd ratings for High Skill videos agreed with true score ratings 76.3% of the time, Faculty ratings agreed 76.6% of the

time, and SP ratings agreed 79.2% of the time, which was not a statistically significant difference in proportions, $p = .69$. For Low Skill videos, Crowd ratings agreed with true score ratings 43.4% of the time, Faculty ratings agreed 60.9% of the time, and SP ratings agreed 51.6% of the time, a statistically significant difference in proportions, $p < .0005$. Post hoc analysis involved pairwise comparisons using the z-test of two proportions with a Bonferroni correction. The proportion of agreement for Faculty raters was statistically significantly higher than for Crowd raters, $p < .05$. The proportions of agreement between SPs and Crowd raters and between SPs and Faculty raters was not statistically significantly different, $p > .05$.

Reliability

As I consider reliability estimates, it is important to reflect upon what "rater" means within the unique rater recruitment strategy of this study. If we imagine ratings laid out in an array of raters as columns and standardized students as rows, moving *down* each column lets us see the same SP or Faculty rater's evaluation of a different student. Similarly, moving *across* each column lets us see each unique rater's rating for the same student. That is, the concept of "rater" takes on a traditional meaning for SP and Faculty raters such that a rater is a single, stable, and unique individual. However, each rating task was posted as a unique assignment (with only the restriction that a Crowd rater could provide ratings once for a given standardized student) in MTurk. This means that, with the exception of those few raters that rated multiple student videos, moving down a column yields a unique rater. This is most akin to the conditions modeled in a one-way random intraclass correlation (ICC (1)) analysis, where individual raters are assumed to be randomly drawn from a larger population of raters. Although this does not preclude me from applying generalizability theory analyses, it does highlight an important point to keep in mind when interpreting the G- and D-studies – that the row array (i.e., the Crowd

ratings of a single standardized student) is arbitrary (in this study it is determined by the order in which ratings for that standardized student came in), and that shuffling this array into another order may change the definition of a "rater" and therefore impact the estimation of variance components.

Rating reliability for ICSF items was analyzed using generalizability theory (Shavelson & Webb, 1991). As described previously, measurement at the item level is of primary interest, because the ICSF is most effectively used to provide detailed feedback for students to improve their performance in the domain represented within each item. As such, items were treated separately. Furthermore, although raters were nested within rater type, the unbalanced number of raters within each rater type necessitated that rater types be treated separately. Treating raters as a facet nested within rater type would have required the exclusion of large amounts of Crowd rater data to balance the number of raters across rater types and estimate variance components. Therefore, rater types were treated separately. This required 42 separate (3 rater types x 14 items) crossed G studies with rater treated as a random facet (student x rater). Variance components for the G study were estimated through the VARCOMP procedure in SPSS Statistics (Version 24) using MINQUE estimation. Table 14 presents variance components for each facet and percentages of total variance, separated by item and rater type. Cells marked by an asterisk represent conditions where variance components were not estimated, because there was little to no variability in ratings for that item within rater type. This condition occurred exclusively in ICSF checklist items and can be interpreted as "perfect" reliability.

Using the variance components derived from the G study, a series of decisions studies (D studies), akin to Spearman-Brown prophecies, were conducted to generate phi coefficients, a reliability coefficient used when one is interested in absolute, rather than relative, performance

levels. Table 15 presents phi coefficients for three conditions (by item and rater type), a single rater, the maximum number of raters of each type available in this study (20 Crowd, 4 SP, 4 Faculty), and one based on cost equivalence.

The cost equivalence phi coefficients were computed in order to determine the reliability one could achieve while holding total spending nearly constant. The single rating cost was $10.74, $6.39, and $3.87 for Faculty, SP, and Crowd raters, respectively. These costs were calculated based on the average Total Task Time *within* rater type and therefore account for the differences in rating speed across rater types. Considering these costs, the number of raters one could recruit with the cash equivalent of a single rating from the most expensive rater would be approximately 1 Faculty rater, 2 SP raters, and 3 Crowd raters for $12. To aid comparison, these phi coefficients are plotted item-by-item in Figures 5 – 18 with a horizontal green bar representing a benchmark of minimal acceptable reliability of .70.

With respect to single-rater reliability for the eight ICSF global items, the highest single rater reliabilities were attributed to Faculty for 7/8 items and SPs for 1/8 items. No single-rater reliabilities for Crowd raters reached the acceptable level of .70 for Crowd raters, while 5/8 Faculty-rated items and 3/8 SP-rated items reached acceptable levels of reliability. With respect to the six checklist items, two items had perfect reliability for SP and Crowd raters while one item had perfect reliability for the Faculty raters. For the remaining four items, Faculty and SPs had the same reliability for one of the items, SPs had two items with the highest single-rater reliabilities, and Faculty had one item with the highest. For Crowd raters, 2/6 checklist items had acceptable levels of reliability, 1/6 for Faculty, and 4/6 for SPs. The checklist results should be interpreted with caution due to low levels of variability in the datasets.

With respect to reliability with the maximum number of raters available in this study for the eight ICSF global items, Faculty raters had the highest reliability for 6/8 items and SPs were tied with Faculty for 2/8 items. With the maximum number of raters, all eight items for all raters had reliabilities above the acceptable level. For the six checklist items, two items had perfect reliability for SP and Crowd raters while one item had perfect reliability for the Faculty raters. For the remaining four items, all rater types were tied for highest reliability for one item, SPs had two items with the highest reliability, and Faculty had one item. All six items for all raters had acceptable levels of reliability, with the exception of one item for SP raters (although this is likely an artifact of low variability in variance component estimation).

With respect to reliability after factoring in cost equivalency for the eight ICSF global items, Faculty had the highest reliability for 2/8 items while SPs had the highest reliability for 6/8 items. For all rater types, 5/8 items met minimum levels of reliability under the cost equivalence conditions. For the six checklist items, two items had perfect reliability for SP and Crowd raters while one item had perfect reliability for the Faculty raters. For the remaining four items, 2/4 had the highest reliability with Crowd raters and 2/4 with SP raters. None of the Faculty-rated items met acceptable levels of reliability, but 3/6 Crowd items and 4/6 SP items did.

A final D-study is presented in Table 16, which outlines the minimum numbers of raters required (by rater type) and the associated costs to achieve minimally acceptable levels of reliability (.70) for each item. For the ICSF global items, it was least expensive to use SPs to reach minimum acceptable reliability for 5/8 items, Faculty for 2/8 items, and Crowd raters for 1/8 items. For the checklist items, 3/6 were least expensive with SPs and 3/6 were least expensive with Crowd raters.

Because the PESC instrument uses a nominal scale, reliability was analyzed using Krippendorff's alpha (kalpha; Hayes & Krippendorff, 2007). Kalpha for multiple raters using a nominal scale requires that the rating matrix be fully crossed. Therefore, because some Crowd ratings were missing or removed from the dataset based on attention check items, several ratings were excluded from reliability analyses to yield a complete matrix with 17 Crowd raters. All four SP and all four Faculty ratings were included in Kalpha analyses. Kalpha coefficeints were computed item-by-item for each rater type using ReCal3 (Freelon, n.d.) and are presented in Table 17. Since Kalpha calculations assume variability in the data, those coefficients marked with an asterisk indicate the inability to generate an estimate due to perfect or near-perfect reliability among raters. For 2/16 items, all raters had perfect reliability, and for 2/16 items Crowd and SP raters had perfect reliability. For the remaining 12 items, 6 items had highest reliabilities associated with Crowd raters, 3 with SPs, and 3 with Faculty. Four Crowd-rated items, 7 SP items, and 5 Faculty items achieved acceptable levels of reliability.

These results should be interpreted with caution due to low variability in ratings. Unfortunately, no procedure currently exists for creating reliability prophecies based on Kalpha coefficients, so it is not possible to use existing reliability levels to prognosticate reliability levels with varying amounts of raters as was possible with the ICSF ratings.

Rater Reactions Questionnaire (RRX)

Likert-type items. Table 18 presents descriptive statistics for all RRX items, separated by Task (ICSF, PESC). Each item from the Rater Reactions Questionnaire (RRX) was analyzed using a 2 x 2 ANOVA. The two factors were the rating Task and Skill Level (High, Low) of the standardized student in the video. While separate videos can be produced for each task (as they were in this study), performance levels of students would be unknown until the ratings actually

took place. As such, it is important to consider the moderating effect of Skill Level to verify that raters perceive task characteristics equivalently across High and Low Skill students. Therefore, when an interaction effect was present, an analysis of simple main effects for Skill Level was performed.

In order to compare initial reactions among raters as well as maintain independence of cases to fit the assumptions of ANOVA analyses, only the first response (in temporal order) for each rater with multiple ratings within a task was retained. After this adjustment, no rater who completed both and ICSF and PESC task remained in the dataset, thereby satisfying the ANOVA assumption of independence.

RRX_1: This item assessed the degree to which Crowd raters felt comfortable evaluating the student's ICS or PE skills. The interaction between Task and Skill Level was statistically significant, $F(1, 143) = 5.72$, $p = .02$, partial $\eta^2 = .04$.

A simple main effects analysis of Skill Level revealed that for ICSF Tasks, mean agreement for High Skill videos was 5.43 ($SD = .50$) and 5.52 ($SD = .59$) for Low Skill videos, a mean difference of .10, 95% CI [-.29, .48], $F(1, 143) = .24$, $p = .63$, partial $\eta^2 = .00$, which was not statistically significant. Thus, ICSF raters viewing High and Low Skill videos did not differ with respect to their comfort evaluating students.

For PESC Tasks, mean agreement for High Skill videos was 5.46 ($SD = .67$) and 4.92 ($SD = 1.23$) for Low Skill videos, a mean difference of .55, 95% CI [.18, .91], $F(1, 143) = 5.73$, $p < .0005$, partial $\eta^2 = .06$, which was statistically significant. Thus, PESC raters viewing Low Skill videos were less comfortable evaluating students than those reviewing High Skill videos.

RRX_2 (ICS only): This item assessed Crowd rater comfort with providing positive written feedback about the student's ICS skills. An independent-samples t-test revealed no differences in comfort providing positive feedback between raters viewing High Skill (M = 5.54, SD = .51) and Low Skill (M = 5.52, SD = .74) videos, M = .01, 95% CI [-.31, .33], $t(68)$ = .07, p = .14.

RRX_3 (ICS only): This item assessed Crowd rater comfort with providing negative written feedback about the student's ICS skills. An independent-samples t-test revealed no differences in comfort providing negative feedback between raters viewing High Skill (M = 5.21, SD = .74) and Low Skill (M = 5.33, SD = .79) videos, M = -.12, 95% CI [-.49, .26], $t(68)$ = -.64, p = .32.

RRX_4: This item measured whether Crowd raters felt they understood the elements of the ICSF or PESC questionnaire well. The interaction between Task and Skill Level was not statistically significant, $F(1, 143) = 3.09$, p = .08, partial η^2 = .02. The main effect of Task was not statistically significant, $F(1, 143) = 1.06$, p = .31, partial η^2 = .01. ICSF raters (M = 5.60) and PESC raters (M = 5.51) understood the elements of the questionnaires similarly well. The main effect of Skill Level was not statistically significant, $F(1, 143) = 3.51$, p = .06, partial η^2 = .02. Raters evaluating High Skill (M = 5.65) and Low Skill (M = 5.46) videos understood the questionnaires similarly well.

RRX_5: This item measured whether Crowd raters felt that the ICSF or PESC questionnaire gave them enough information to make decisions about the student's ICS or PE skills. The interaction between Task and Skill Level was not statistically significant, $F(1, 143)$ = 3.00, p = .09, partial η^2 = .02.

The main effect of Task was not statistically significant, $F(1, 143) = 1.60$, $p = .21$, partial $\eta^2 = .01$. ICSF raters ($M = 5.40$) and PESC raters ($M = 5.27$) had similar levels of agreement about whether the rating tool gave them enough information to make decisions about the student's performance.

The main effect of Skill Level was statistically significant such that raters reviewing High Skill Videos ($M = 5.54$) agreed more than raters reviewing Low Skill videos ($M = 5.15$) that they had enough information to make decisions about a student's performance, $F(1, 143) = 9.10$, $p < .0005$, partial $\eta^2 = .06$.

RRX_6: This item measured whether Crowd raters felt the instructions for completing the PESC or ICSF were clear. The interaction between Task and Skill Level was statistically significant, $F(1, 143) = 4.49$, $p = .04$, partial $\eta 2 = .03$.

A simple main effects analysis of Skill Level revealed that for ICSF Tasks, mean agreement for High Skill videos was 5.68 ($SD = .48$) and 5.57 ($SD = .55$) for Low Skill videos, a mean difference of .11, 95% CI [-.16, 0.38], $F(1, 143) = 0.62$, $p = .43$, partial $\eta^2 = .00$, which was not statistically significant. Thus, ICSF raters viewing High and Low Skill videos did not differ with respect to their perceptions of instruction clarity.

For PESC Tasks, mean agreement for High Skill videos was 5.78 ($SD = .42$) and 5.28 ($SD = 0.74$) for Low Skill videos, a mean difference of .50, 95% CI [.25, .76], $F(1, 143) = 15.52$, $p < .0005$, partial $\eta^2 = .10$, which was statistically significant. Thus, PESC raters viewing Low Skill videos perceived the task instructions to be less clear than those reviewing High Skill videos.

RRX_7: This item measured Crowd rater motivation to make accurate ratings of the student's ICS or PE skills. The interaction between Task and Skill Level was statistically significant, $F(1, 143) = 5.33$, $p = .02$, partial $\eta^2 = .04$.

A simple main effects analysis of Skill Level revealed that for ICSF Tasks, mean agreement for High Skill videos was 5.36 (*SD* = .99) and 5.71 (*SD* = .60) for Low Skill videos, a mean difference of .36, 95% CI [.08, .64], $F(1, 143) = 6.28$, $p = .01$, partial $\eta^2 = .00$, which was statistically significant. Thus, ICSF raters viewing High Skill videos reported a lower level of motivation to make accurate ratings than those rating Low Skill videos.

For PESC Tasks, mean agreement for High Skill videos was 5.93 (*SD* = .26) and 5.83 (*SD* = .38) for Low Skill videos, a mean difference of .09, 95% CI [-.17, .36], $F(1, 143) = .49$, $p = .49$, partial $\eta^2 = .00$, which was not statistically significant. Thus, PESC raters viewing Low and High Skill videos reported similar levels of motivation in evaluating student performance accurately.

RRX_8: This item measured Crowd rater motivation to be fair in evaluating the student's ICS or PE skills. The interaction between Task and Skill Level was statistically significant, $F(1, 143) = 4.83$, $p = .03$, partial $\eta^2 = .03$.

A simple main effects analysis of Skill Level revealed that for ICSF Tasks, mean agreement for High Skill videos was 5.54 (*SD* = .51) and 5.79 (*SD* = .47) for Low Skill videos, a mean difference of .25, 95% CI [.01, .49], $F(1, 143) = 4.28$, $p = .04$, partial $\eta^2 = .03$, which was statistically significant. Thus, ICSF raters viewing High Skill videos reported a lower level of motivation to be fair in their ratings than those rating Low Skill videos.

For PESC Tasks, mean agreement for High Skill videos was 5.78 (*SD* = .26) and 5.67 (*SD* = .59) for Low Skill videos, a mean difference of .11, 95% CI [-.11, .34], $F(1, 143) =$

1.01, p = .32, partial η^2 = .01, which was not statistically significant. Thus, PESC raters viewing Low and High Skill videos reported similar levels of motivation to be fair in evaluating student performance.

RRX_9: This item measured the degree to which Crowd raters felt that patients, like themselves, should be able to provide feedback about the ICS or PE skills of future physicians. The interaction between Task and Skill Level was statistically significant, $F(1, 143) = 6.97$, p = .01, partial η^2 = .05.

A simple main effects analysis of Skill Level revealed that for ICSF Tasks, mean agreement for High Skill videos was 5.43 (SD = .63) and 5.60 (SD = .59) for Low Skill videos, a mean difference of .17, 95% CI [-.18, .51], $F(1, 143) = .92$, p = .34, partial η^2 = .01, which was not statistically significant. Thus, ICSF raters viewing High and Low Skill videos had similar levels of agreement that raters should be able to provide feedback about interpersonal and communication skills.

For PESC Tasks, mean agreement for High Skill videos was 5.49 (SD = .71) and 5.03 (SD = .88) for Low Skill videos, a mean difference of .46, 95% CI [.14, .78], $F(1, 143) = 8.04$, p = .01, partial η^2 = .05, which was statistically significant. Thus, PESC raters viewing High Skill videos agreed more than raters viewing Low Skill videos that they should be able to provide feedback about physical exam skills.

RRX_10: This item measured the degree to which Crowd raters felt that patient, like themselves, are capable of providing feedback about the ICS or PE skills of future physicians. The interaction between Task and Skill Level was statistically significant, $F(1, 143) = 6.13$, p = .02, partial η^2 = .04.

A simple main effects analysis of Skill Level revealed that for ICSF Tasks, mean agreement for High Skill videos was 5.29 (*SD* = 0.90) and 5.48 (*SD* = 0.63) for Low Skill videos, a mean difference of .19, 95% CI [-.23, .61], $F(1, 143) = .81$, $p = .37$, partial $\eta^2 = .01$, which was not statistically significant. Thus, ICSF raters viewing High and Low Skill videos had similar levels of agreement that patients are capable of providing feedback about interpersonal and communication skills.

For PESC Tasks, mean agreement for High Skill videos was 5.39 (*SD* = .77) and 4.86 (*SD* = 1.15) for Low Skill videos, a mean difference of .53, 95% CI [.14, .92], $F(1, 143) = 7.09$, $p = .01$, partial $\eta^2 = .05$, which was statistically significant. Thus, PESC raters viewing High Skill videos agreed more than raters viewing Low Skill videos that patients are capable of providing feedback about physical exam skills.

RRX_11: The item measured the degree to which Crowd raters felt that they need additional training to make ICSF or PE evaluations in the future. The interaction between Task and Skill Level was not statistically significant, $F(1, 143) = .01$, $p = .95$, partial $\eta^2 = .00$.

The main effect of Task was statistically significant, $F(1, 143) = 9.24$, $p < .0005$, partial $\eta^2 = .06$. PESC raters (*M* = 2.74, *SD* = .16) felt that they need additional training more than ICSF raters (*M* = 2.17, *SD* = 1.12).

The main effect of Skill Level was statistically significant, $F(1, 143) = 2.74$, $p = .03$, partial $\eta^2 = .03$. Raters evaluating Low Skill (*M* = 2.64, *SD* = 1.35) videos felt that they needed additional training more than those reviewing High Skill videos (*M* = 2.28, *SD* = 1.17).

Open-ended items. A quantitative content analysis was used to develop themes found in the open-ended Crowd rater comments. Because there was little prior knowledge about Crowd rater reactions to rating medical student performance, I used an inductive approach with iterative

open coding and a phrase as the unit of analysis. Phrases were coded using QDA Miner Lite. Comments were aggregated across videos within rating task for any raters who participated in a task more than once. Although this caused some duplicate phrases for individual raters, results of this content analysis are presented as the percentage raters that made a comment within a particular category. Therefore, duplicate comments within rater do not impact the data.

After reading the text of the PE and ICS task comments multiple times, the following categories emerged across both sets of comments:

- Task Positive: Positive comments about procedural elements of the task (e.g., clarity of instructions)
- Task Negative: Negative comments about procedural elements of the task (e.g., layout of the ICSF form)
- Technical Concerns: Comments about elements of the task not functioning (e.g., video not playing immediately)
- Comfortable Evaluating: Comments expressing the rater's confidence about evaluating the student
- Apprehensive Evaluating: Comments expressing the rater's apprehension about evaluating the student
- Face Blurring: Comments about the face blurring feature making the task more difficult
- Student Performance: Comments about the specific student's performance rather than rater's experience of the task
- Objective: Comments about the rater trying his or her best to be objective or fair
- Enjoy: Comments about the rater enjoying or liking the task or having fun
- Interesting: Comments about the task being interesting or engaging

- Patient voice: Comments supporting rater perceptions that patients should have a voice in contributing to the training of physicians or evaluating physicians in training.

Percentages of raters with comments in each category are presented in Table 19. All comments by category for the ICS task are presented in Appendix 15. All comments by category for the PE task are presented in Appendix 16.

Table 1. Video and task properties.

Variable	Video					
Standardized Student	BK	CE	CS	DB	EF	JJ
Standardized Patient	RL	MH	MH	RL	RL	MH
	ICS Task					
Skill Level	Low	Low	Low	High	Low	High
Video Length (mm:ss)	9:14	6:17	5:20	12:51	6:06	9:43
Task Length (mm)	29	26	25	33	26	30
Rater Fee/Rating	$4.35	$3.90	$3.75	$4.95	$3.90	$4.50
Mturk Fee/Rating	$1.74	$1.56	$1.50	$1.98	$1.56	$1.80
Total Fee/Rating	$6.09	$5.46	$5.25	$6.93	$5.46	$6.30
Ratings Collected	20	20	20	20	20	20
Ratings Excluded*	2	0	0	1	4	0
Ratings Analyzed	18	20	20	19	16	20
	PE Task					
Skill Level	High	Low	Low	High	Low	High
Video Length (mm:ss)	High	Low	Low	High	Low	High
Task Length (mm)	3:16	2:20	1:42	2:59	2:31	3:00
Rater Fee/Rating	$2.25	$2.25	$2.25	$2.25	$2.25	$2.25
Mturk Fee/Rating	$0.90	$0.90	$0.90	$0.90	$0.90	$0.90
Total Fee/Rating	$3.15	$3.15	$3.15	$3.15	$3.15	$3.15
Ratings Collected	20	20	20	19	20	20
Ratings Excluded*	1	0	3	2	3	3
Ratings Analyzed	19	20	17	17	17	17

* Based on attention check

Table 2. Rater schematic.

Video	True Score	Standardized Patient				Faculty				Crowd Rater				
----------	Sim Manager	SP 1	SP 2	SP 3	SP 4	FA 1	FA 2	FA 3	FA 4	CR 1	CR 2	CR 3	….	CR 20
BKRL														
CEMH														
CSMH														
DBRL														
EFRL														
JJMH														

Table 3. Comparison of Crowd rater and U.S. demographics.

Gender	**Crowd**	**U.S.***
Male	42.7%	49.2%
Female	57.3%	50.8%
Age		
20-24	8.70%	9.4%
25-34	39.30%	18.4%
35-44	29.30%	17.1%
45-54	12.70%	18.0%
55-64	8.00%	17.1%
65+	2.00%	20.0%
Education**		
Less than High School	1.53%	12.9%
High School Graduate	13.74%	27.6%
Some college, no degree	22.90%	20.7%
Associate's Degree	15.27%	8.2%
Bachelor's Degree	35.11%	19.0%
Graduate/Professional Degree	11.45%	11.6%
Marital Status		
Married	36.0%	47.5%
Divorced	8.0%	11.0%
Separated	2.00%	2.1%
Single/Never Married	54.00%	33.5%
Race		
White	77.30%	75.8%
Black/African American	9.33%	13.9%
Hispanic/Latino	6.00%	***
American Indian/Alaska Native	1.33%	1.7%
Asian	4.0%	6.4%
Other Race/Combination	2.0%	5.3%
Income		
Less than 25,000	24.0%	22.0%
$25,000 to $34,999	12.7%	9.8%
$35,000 to $49,999	22.7%	13.2%
$50,000 to $74,999	21.3%	17.8%
$75,000 to $99,999	12.7%	12.2%
$100,000 to $149,999	5.3%	13.6%
$150,000 or more	1.3%	11.3%

* Derived from the U.S. Census Bureau (2015). *American Community Survey 1-year estimates.*
** For participants 25 and older
*** American Community Survey does not classify Hispanic/Latino as a race

Table 4. Task timing (video x task x rater type).

Video	Task	Video Length	Crowd k*	Mean Total Task Time			Mean Instrument Time		
				Crowd (SD)	SP	Faculty	Crowd	SP	Faculty
BKRL	ICS	9:14	9	20:34 (5:58)	15:45	13:52	11:20	6:31	4:38
	PE	3:16	11	16:29 (6:18)	5:25	3:25	13:13	2:09	0:09
CEMH	ICS	6:17	15	21:00 (9:09)	20:21	8:10	14:43	14:04	1:53
	PE	2:20	13	14:04 (3:02)	5:20	6:35	11:44	3:00	4:15
CSMH	ICS	5:20	7	17:01 (10:32)	15:37	7:23	11:41	10:17	2:03
	PE	1:42	10	14:26 (7:21)	5:18	4:56	12:44	3:36	3:14
DBRL	ICS	12:51	17	24:24 (10:01)	14:00	12:55	11:33	2:51	0:04
	PE	2:59	14	15:00 (5:20)	4:01	9:04	12:01	1:02	6:05
EFRL	ICS	6:06	10	20:33 (9:55)	19:47	8:32	14:27	13:41	2:26
	PE	2:31	11	17:13 (4:38)	5:45	6:28	14:42	3:14	3:57
JJMH	ICS	9:43	6	25:30 (2:51)	18:07	11:08	15:47	8:24	1:25
	PE	3:00	12	15:26 (4:19)	6:27	5:59	12:26	3:27	2:59
Mean	**ICS**	**8:15**	--	**21:30**	**17:16**	**10:20**	**13:15**	**9:18**	**2:05**
	PE	**2:38**	--	**15:26**	**5:22**	**6:04**	**12:48**	**2:44**	**3:26**

* Timing outliers removed

Table 5. Timing to complete rating packages (video x task x rater type).

Video	Task	Crowd*	SP	Faculty
BKRL	ICS	00:3:29	SP** for ICS and PE = 10:04:52	BhvFac** for ICS = 16:12:29
	PE	00:1:52		
CEMH	ICS	00:2:58		
	PE	00:1:49		
CSMH	ICS	00:3:49		
	PE	00:2:35		
DBRL	ICS	00:3:31		PhysFac** for PE = 8:18:56
	PE	00:3:33		
EFRL	ICS	00:5:43		
	PE	00:1:32		
JJMH	ICS	00:3:08		
	PE	00:3:12		

Note: All timing data is presented as DD:HH:MM
* Each package contained 20 ratings for 6 students. All HIT assignments within task (ICS, PE) were deployed in parallel, but the tasks were deployed at separate times.
** 4 ratings for 6 students

Table 6. Response distribution for ICSF checklist items (item x rater type x video).

			Video											
			BKRL		CEMH		CSMH		DBRL		EFRL		JJMH	
Item	**Rater**	**Metric**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
ICS_1	Crowd	%	6%	94%	5%	95%		100%	5%	95%		100%	5%	95%
		Count	1	17	1	19		20	1	18		16	1	19
	Faculty	%		100%		100%		100%		100%		100%		100%
		Count		4		4		4		4		4		4
	SP	%		100%		100%		100%		100%		100%		100%
		Count		4		4		4		4		4		4
ICS_2	Crowd	%		100%	5%	95%		100%		100%		100%		100%
		Count		18	1	19		20		19		16		20
	Faculty	%		100%	50%	50%		100%		100%	75%	25%		100%
		Count		4	2	2		4		4	3	1		4
	SP	%		100%		100%		100%		100%		100%		100%
		Count		4		4		4		4		4		4
ICS_3	Crowd	%		100%	40%	60%		100%	5%	95%	69%	31%		100%
		Count		18	8	12		20	1	18	11	5		20
	Faculty	%		100%	75%	25%		100%		100%	50%	50%		100%
		Count		4	3	1		4		4	2	2		4
	SP	%		100%	50%	50%		100%		100%	75%	25%		100%
		Count		4	2	2		4		4	3	1		4
ICS_10	Crowd	%		100%	60%	40%	15%	85%		100%	13%	88%		100%
		Count		18	12	8	3	17		19	2	14		20
	Faculty	%	25%	75%	75%	25%	75%	25%		100%	25%	75%	25%	75%
		Count	1	3	3	1	3	1		4	1	3	1	3
	SP	%		100%	100%		25%	75%		100%		100%		100%
		Count		4	4		1	3		4		4		4
ICS_11	Crowd	%		100%	65%	35%	25%	75%		100%	19%	81%		100%
		Count		18	13	7	5	15		19	3	13		20
	Faculty	%	25%	75%	100%		100%		25%	75%	100%		50%	50%
		Count	1	3	4		4		1	3	4		2	2
	SP	%		100%	100%		100%			100%	75%	25%		100%
		Count		4	4		4			4	3	1		4
ICS_12	Crowd	%	17%	83%	55%	45%	15%	85%	11%	89%	88%	13%	10%	90%
		Count	3	15	11	9	3	17	2	17	14	2	2	18
	Faculty	%		100%	75%	25%	50%	50%		100%	75%	25%		100%
		Count		4	3	1	2	2		4	3	1		4
	SP	%	50%	50%	50%	50%	50%	50%		100%	75%	25%	50%	50%
		Count	2	2	2	2	2	2		4	3	1	2	2

Table 7. Response distribution for ICSF global items (item x rater type x video).

			Video														
			BKRL					CEMH					CSMH				
Item	**Rater**	**Metric**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
ICS_4	Crowd	%			17%	39%	44%		5%	45%	15%	35%		5%	25%	50%	20%
		Count			3	7	8		1	9	3	7		1	5	10	4
	Faculty	%			75%	25%		25%	50%		25%		25%	25%	50%		
		Count			3	1		1	2		1		1	1	2		
	SP	%			50%	50%				100%				25%	50%	25%	
		Count			2	2				4				1	2	1	
ICS_5	Crowd	%			17%	44%	39%		20%	30%	35%	15%		20%	30%	30%	20%
		Count			3	8	7		4	6	7	3		4	6	6	4
	Faculty	%			50%	50%			75%	25%			25%	50%	25%		
		Count			2	2			3	1			1	2	1		
	SP	%			50%	50%			50%	25%	25%			25%	50%	25%	
		Count			2	2			2	1	1			1	2	1	
ICS_6	Crowd	%			17%	50%	33%	15%	45%	10%	5%	25%		15%	35%	45%	5%
		Count			3	9	6	3	9	2	1	5		3	7	9	1
	Faculty	%		25%	25%	50%		100%						100%			
		Count		1	1	2		4						4			
	SP	%			75%	25%			50%	50%				25%	75%		
		Count			3	1			2	2				1	3		
ICS_7	Crowd	%	6%	6%	6%	39%	44%	10%	10%	25%	30%	25%		20%	25%	35%	20%
		Count	1	1	1	7	8	2	2	5	6	5		4	5	7	4
	Faculty	%		25%	25%	50%			75%	25%				75%	25%		
		Count		1	1	2			3	1				3	1		
	SP	%			75%	25%				75%	25%			25%	50%	25%	
		Count			3	1				3	1			1	2	1	
ICS_8	Crowd	%		6%	22%	56%	17%	60%	10%	25%		5%		5%	40%	50%	5%
		Count		1	4	10	3	12	2	5		1		1	8	10	1
	Faculty	%		50%	50%			100%					50%	25%	25%		
		Count		2	2			4					2	1	1		
	SP	%	25%		50%	25%		50%	50%						25%	75%	
		Count	1		2	1		2	2						1	3	
ICS_9	Crowd	%			6%	44%	50%	5%	20%	15%	40%	20%	5%	5%	35%	35%	20%
		Count			1	8	9	1	4	3	8	4	1	1	7	7	4
	Faculty	%		25%		75%		25%	50%	25%				75%	25%		
		Count		1		3		1	2	1				3	1		
	SP	%			50%	50%			50%	25%	25%				100%		
		Count			2	2			2	1	1				4		

Table 7 (Continued)

			Video														
			BKRL					CEMH					CSMH				
Item	**Rater**	**Metric**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
ICS_13	Crowd	%		11%	11%	50%	28%	10%	10%	45%	20%	15%		15%	30%	35%	20%
		Count		2	2	9	5	2	2	9	4	3		3	6	7	4
	Faculty	%			75%	25%		50%	25%	25%			25%	75%			
		Count			3	1		2	1	1			1	3			
	SP	%			75%	25%			25%	75%				25%	50%	25%	
		Count			3	1			1	3				1	2	1	
ICS_14	Crowd	%		6%	28%	22%	44%	10%	35%	40%	10%	5%	5%	25%	25%	10%	35%
		Count		1	5	4	8	2	7	8	2	1	1	5	5	2	7
	Faculty	%	25%		75%			100%					75%	25%			
		Count	1		3			4					3	1			
	SP	%			75%	25%			50%	50%				25%	75%		
		Count			3	1			2	2				1	3		

Table 7 (Continued)

			Video														
			DBRL					EFRL					JJMH				
Item	**Rater**	**Metric**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
ICS_4	Crowd	%			5%	42%	53%		6%	50%	38%	6%			15%	10%	75%
		Count			1	8	10		1	8	6	1			3	2	15
	Faculty	%			25%	75%		25%	50%	25%					25%	75%	
		Count			1	3		1	2	1					1	3	
	SP	%				75%	25%	25%		75%					25%	50%	25%
		Count				3	1	1		3					1	2	1
ICS_5	Crowd	%			16%	42%	42%		44%	38%	13%	6%		5%	5%	30%	60%
		Count			3	8	8		7	6	2	1		1	1	6	12
	Faculty	%		25%		50%	25%	75%	25%						25%	75%	
		Count		1		2	1	3	1						1	3	
	SP	%				50%	50%		75%	25%					25%	75%	
		Count				2	2		3	1					1	3	
ICS_6	Crowd	%			5%	53%	42%	13%	50%	38%						40%	60%
		Count			1	10	8	2	8	6						8	12
	Faculty	%			25%	75%		100%							50%	25%	25%
		Count			1	3		4							2	1	1
	SP	%				50%	50%	50%	25%	25%						50%	50%
		Count				2	2	2	1	1						2	2
ICS_7	Crowd	%		5%	5%	11%	79%	6%	19%	63%	13%					30%	70%
		Count		1	1	2	15	1	3	10	2					6	14
	Faculty	%				75%	25%	25%	50%	25%						100%	
		Count				3	1	1	2	1						4	
	SP	%				25%	75%		25%	50%	25%					75%	25%
		Count				1	3		1	2	1					3	1
ICS_8	Crowd	%		5%	11%	37%	47%	63%	19%	19%				5%	15%	30%	50%
		Count		1	2	7	9	10	3	3				1	3	6	10
	Faculty	%				50%	50%	100%						50%		50%	
		Count				2	2	4						2		2	
	SP	%					100%	75%	25%							50%	50%
		Count					4	3	1							2	2
ICS_9	Crowd	%			5%	32%	63%	19%	38%	19%	25%				10%	20%	70%
		Count			1	6	12	3	6	3	4				2	4	14
	Faculty	%			25%	75%		100%								100%	
		Count			1	3		4								4	
	SP	%				75%	25%		25%	50%	25%					75%	25%
		Count				3	1		1	2	1					3	1

Table 7 (Continued)

			Video														
			DBRL					EFRL					JJMH				
Item	**Rater**	**Metric**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
ICS_13	Crowd	%			16%	21%	63%	25%	31%	38%	6%				5%	20%	75%
		Count			3	4	12	4	5	6	1				1	4	15
	Faculty	%				100%		25%	50%	25%					50%	25%	25%
		Count				4		1	2	1					2	1	1
	SP	%				50%	50%		50%	50%						75%	25%
		Count				2	2		2	2						3	1
ICS_14	Crowd	%	5%		11%	32%	53%	31%	38%	31%					10%	25%	65%
		Count	1		2	6	10	5	6	5					2	5	13
	Faculty	%		25%		75%		100%							50%	50%	
		Count		1		3		4							2	2	
	SP	%				25%	75%	25%	25%	50%						75%	25%
		Count				1	3	1	1	2						3	1

Table 8. Descriptive statistics for ICSF (item x rater type x video).

Item	Rater		Video					
			BKRL	CEMH	CSMH	DBRL	EFRL	JJMH
		Mean	0.94	0.95	1.00	0.95	1.00	0.95
	Crowd	N	18	20	20	19	16	20
		SD	0.24	0.22	0.00	0.23	0.00	0.22
		Mean	1.00	1.00	1.00	1.00	1.00	1.00
ICS_1_chx	Faculty	N	4	4	4	4	4	4
		SD	0.00	0.00	0.00	0.00	0.00	0.00
		Mean	1.00	1.00	1.00	1.00	1.00	1.00
	SP	N	4	4	4	4	4	4
		SD	0.00	0.00	0.00	0.00	0.00	0.00
		Mean	1.00	0.95	1.00	1.00	1.00	1.00
	Crowd	N	18	20	20	19	16	20
		SD	0.00	0.22	0.00	0.00	0.00	0.00
		Mean	1.00	0.50	1.00	1.00	0.25	1.00
ICS_2_chx	Faculty	N	4	4	4	4	4	4
		SD	0.00	0.58	0.00	0.00	0.50	0.00
		Mean	1.00	1.00	1.00	1.00	1.00	1.00
	SP	N	4	4	4	4	4	4
		SD	0.00	0.00	0.00	0.00	0.00	0.00
		Mean	1.00	0.60	1.00	0.95	0.31	1.00
	Crowd	N	18	20	20	19	16	20
		SD	0.00	0.50	0.00	0.23	0.48	0.00
		Mean	1.00	0.25	1.00	1.00	0.50	1.00
ICS_3_chx	Faculty	N	4	4	4	4	4	4
		SD	0.00	0.50	0.00	0.00	0.58	0.00
		Mean	1.00	0.50	1.00	1.00	0.25	1.00
	SP	N	4	4	4	4	4	4
		SD	0.00	0.58	0.00	0.00	0.50	0.00
		Mean	4.28	3.80	3.85	4.47	3.44	4.60
	Crowd	N	18	20	20	19	16	20
		SD	0.75	1.01	0.81	0.61	0.73	0.75
		Mean	3.25	2.25	2.25	3.75	2.00	3.75
ICS_4	Faculty	N	4	4	4	4	4	4
		SD	0.50	1.26	0.96	0.50	0.82	0.50
		Mean	3.50	3.00	3.00	4.25	2.50	4.00
	SP	N	4	4	4	4	4	4
		SD	0.58	0.00	0.82	0.50	1.00	0.82
		Mean	4.22	3.45	3.50	4.26	2.81	4.45
	Crowd	N	18	20	20	19	16	20
		SD	0.73	1.00	1.05	0.73	0.91	0.83
		Mean	3.50	2.25	2.00	3.75	1.25	3.75
ICS_5	Faculty	N	4	4	4	4	4	4
		SD	0.58	0.50	0.82	1.26	0.50	0.50
		Mean	3.50	2.75	3.00	4.50	2.25	3.75
	SP	N	4	4	4	4	4	4
		SD	0.58	0.96	0.82	0.58	0.50	0.50
		Mean	4.17	2.80	3.40	4.37	2.25	4.60
	Crowd	N	18	20	20	19	16	20
		SD	0.71	1.47	0.82	0.60	0.68	0.50
		Mean	3.25	1.00	2.00	3.75	1.00	3.75
ICS_6	Faculty	N	4	4	4	4	4	4
		SD	0.96	0.00	0.00	0.50	0.00	0.96
		Mean	3.25	2.50	2.75	4.50	1.75	4.50
	SP	N	4.00	4.00	4.00	4.00	4.00	4.00
		SD	0.50	0.58	0.50	0.58	0.96	0.58

Table 8 (Continued)

Item	Rater Type		Video BKRL	CEMH	CSMH	DBRL	EFRL	JJMH
		Mean	4.11	3.50	3.55	4.63	2.81	4.70
	Crowd	N	18	20	20	19	16	20
		SD	1.13	1.28	1.05	0.83	0.75	0.47
		Mean	3.25	2.25	2.25	4.25	2.00	4.00
ICS_7	Faculty	N	4	4	4	4	4	4
		SD	0.96	0.50	0.50	0.50	0.82	0.00
		Mean	3.25	3.25	3.00	4.75	3.00	4.25
	SP	N	4	4	4	4	4	4
		SD	0.50	0.50	0.82	0.50	0.82	0.50
		Mean	3.83	1.80	3.55	4.26	1.56	4.25
	Crowd	N	18	20	20	19	16	20
		SD	0.79	1.15	0.69	0.87	0.81	0.91
		Mean	2.50	1.00	1.75	4.50	1.00	3.00
ICS_8	Faculty	N	4	4	4	4	4	4
		SD	0.58	0.00	0.96	0.58	0.00	1.15
		Mean	2.75	1.50	3.75	5.00	1.25	4.50
	SP	N	4	4	4	4	4	4
		SD	1.26	0.58	0.50	0.00	0.50	0.58
		Mean	4.44	3.50	3.60	4.58	2.50	4.60
	Crowd	N	18	20	20	19	16	20
		SD	0.62	1.19	1.05	0.61	1.10	0.68
		Mean	3.50	2.00	2.25	3.75	1.00	4.00
ICS_9	Faculty	N	4	4	4	4	4	4
		SD	1.00	0.82	0.50	0.50	0.00	0.00
		Mean	3.50	2.75	3.00	4.25	3.00	4.25
	SP	N	4	4	4	4	4	4
		SD	0.58	0.96	0.00	0.50	0.82	0.50
		Mean	1.00	0.40	0.85	1.00	0.88	1.00
	Crowd	N	18	20	20	19	16	20
		SD	0.00	0.50	0.37	0.00	0.34	0.00
		Mean	0.75	0.25	0.25	1.00	0.75	0.75
ICS_10_chx	Faculty	N	4	4	4	4	4	4
		SD	0.50	0.50	0.50	0.00	0.50	0.50
		Mean	1.00	0.00	0.75	1.00	1.00	1.00
	SP	N	4	4	4	4	4	4
		SD	0.00	0.00	0.50	0.00	0.00	0.00
		Mean	1.00	0.35	0.75	1.00	0.81	1.00
	Crowd	N	18	20	20	19	16	20
		SD	0.00	0.49	0.44	0.00	0.40	0.00
		Mean	0.75	0.00	0.00	0.75	0.00	0.50
ICS_11_chx	Faculty	N	4	4	4	4	4	4
		SD	0.50	0.00	0.00	0.50	0.00	0.58
		Mean	1.00	0.00	0.00	1.00	0.25	1.00
	SP	N	4	4	4	4	4	4
		SD	0.00	0.00	0.00	0.00	0.50	0.00
		Mean	0.83	0.45	0.85	0.89	0.13	0.90
	Crowd	N	18	20	20	19	16	20
		SD	0.38	0.51	0.37	0.32	0.34	0.31
		Mean	1.00	0.25	0.50	1.00	0.25	1.00
ICS_12_chx	Faculty	N	4	4	4	4	4	4
		SD	0.00	0.50	0.58	0.00	0.50	0.00
		Mean	0.50	0.50	0.50	1.00	0.25	0.50
	SP	N	4	4	4	4	4	4
		SD	0.58	0.58	0.58	0.00	0.50	0.58

Table 8 (Continued)

Item	Rater Type		Video BKRL	CEMH	CSMH	DBRL	EFRL	JJMH
		Mean	3.94	3.20	3.60	4.47	2.25	4.70
	Crowd	N	18	20	20	19	16	20
		SD	0.94	1.15	0.99	0.77	0.93	0.57
		Mean	3.25	1.75	1.75	4.00	2.00	3.75
ICS_13	Faculty	N	4	4	4	4	4	4
		SD	0.50	0.96	0.50	0.00	0.82	0.96
		Mean	3.25	2.75	3.00	4.50	2.50	4.25
	SP	N	4	4	4	4	4	4
		SD	0.50	0.50	0.82	0.58	0.58	0.50
		Mean	4.06	2.65	3.45	4.26	2.00	4.55
	Crowd	N	18	20	20	19	16	20
		SD	1.00	0.99	1.36	1.05	0.82	0.69
		Mean	2.50	1.00	1.25	3.50	1.00	3.50
ICS_14	Faculty	N	4	4	4	4	4	4
		SD	1.00	0.00	0.50	1.00	0.00	0.58
		Mean	3.25	2.50	2.75	4.75	2.25	4.25
	SP	N	4	4	4	4	4	4
		SD	0.50	0.58	0.50	0.50	0.96	0.50
		Mean	38.83	28.40	33.95	41.11	23.75	42.30
	Crowd	N	18	20	20	19	16	20
		SD	4.99	8.06	6.99	3.77	5.00	4.07
		Mean	30.50	15.75	19.25	37.00	14.00	34.75
ICS_Total	Faculty	N	4	4	4	4	4	4
		SD	2.08	4.19	3.59	4.24	3.16	4.99
		Mean	31.75	24.00	28.50	42.50	22.25	39.25
	SP	N	4	4	4	4	4	4
		SD	4.03	4.24	3.70	2.89	4.57	2.63

Table 9. Response distribution for PESC (item x rater type x video).

		Video																					
		BKRL				CEMH				CSMH				DBRL				EFRL				JJMH			
Item	Rater	C	I	N	O	C	I	N	O	C	I	N	O	C	I	N	O	C	I	N	O	C	I	N	O
	Crowd	.63			.37	.90			.10	.88		.06	.06	.47		.29	.24	.82			.18	.53		.06	.41
PE_CVR_1	Faculty	.50			.50	1.00				1.00				.50			.50	1.00				1.00			
	SP	1.00				1.00				1.00				.75		.25		1.00				1.00			
	Crowd	1.00				.65	.15	.20		.65	.18	.18		1.00				.18	.06	.76		1.00			
PE_CVR_2	Faculty	1.00				.75		.25		.50	.25	.25		1.00						1.00		1.00			
	SP	1.00				.75		.25		.50	.50			1.00						1.00		1.00			
	Crowd	1.00				.90	.10			.88	.06	.06		1.00				.12		.88		1.00			
PE_CVR_3	Faculty	1.00				.75	.25			.50	.50			1.00						1.00		1.00			
	SP	1.00				.75	.25			.75	.25			1.00						1.00		.75	.25		
	Crowd	1.00				.90	.10			.88	.06	.06		1.00				.12		.88		1.00			
PE_CVR_4	Faculty	1.00				.75	.25			.50	.50			1.00						1.00		.75	.25		
	SP	1.00				.75	.25			.75	.25			1.00						1.00		.75	.25		
	Crowd	1.00				.55	.25	.20		.24	.35	.41		1.00				.06		.88	.06	1.00			
PE_CVR_5	Faculty	1.00					.75	.25			.75	.25		1.00						1.00		1.00			
	SP	1.00				.25	.75			.25	.75			1.00						1.00		1.00			
	Crowd	.95	.05			.90	.10			.88	.06	.06		.94	.06			.18		.82		1.00			
PE_CVR_6	Faculty	1.00				.50	.50			.75	.25			1.00						1.00		1.00			
	SP	1.00				.75	.25			.75	.25			1.00						1.00		.75	.25		
	Crowd	1.00				.95	.05			.94	.06			.94			.06	.53		.06	.41	1.00			
PE_CVR_7	Faculty	1.00				.75	.25			.75	.25			1.00						.25	.75	1.00			
	SP	1.00				.75	.25			.75	.25			1.00						.50	.50	.75	.25		
	Crowd	1.00				.95	.05			1.00				.94			.06	.53		.18	.29	1.00			
PE_CVR_8	Faculty	1.00				.75	.25			.75	.25			1.00						.25	.75	.75	.25		
	SP	1.00				.75	.25			.75	.25			1.00						.50	.50	.75	.25		
	Crowd	1.00				.85	.15			1.00				.94	.06			.71		.06	.24	1.00			
PE_CVR_9	Faculty	1.00				.75	.25			1.00				1.00							1.00	1.00			
	SP	1.00				1.00				1.00				1.00				.75			.25	1.00			
	Crowd	.89	.11			.90	.10			.94		.06		1.00				.65		.12	.24	1.00			
PE_CVR_10	Faculty	1.00				.50	.50			1.00				1.00							1.00	1.00			
	SP	.75	.25			.75	.25			.75	.25			1.00				.75			.25	.75	.25		
	Crowd	1.00				.95	.05			.88	.06	.06		1.00				.18		.71	.12	1.00			
PE_CVR_11	Faculty	1.00				.50			.50	1.00				1.00						.50	.50	.75			.25
	SP	1.00				.75	.25			.75	.25			1.00						.75	.25	.75	.25		

Note: Response distributions are presented as proportions within video and within rater type.
C = Done Correctly
I = Done Incorrectly
N = Not Done
O = Observation Obscured

Table 9 (Continued)

		Video																					
		BKRL				CEMH				CSMH				DBRL				EFRL				JJMH			
Item	Rater	C	I	N	O	C	I	N	O	C	I	N	O	C	I	N	O	C	I	N	O	C	I	N	O
	Crowd	1.00				.95	.05			.88	.06	.06		1.00				.24		.65	.12	1.00			
PE_CVR_12	Faculty	1.00				.50			.50	1.00				1.00						.50	.50	.75			.25
	SP	1.00				.75	.25			.75	.25			1.00						.75	.25	.75	.25		
	Crowd	1.00				.90	.05	.05		.88	.06	.06		.94	.06			.18		.71	.12	1.00			
PE_CVR_13	Faculty	1.00				.50			.50	1.00				1.00						.50	.50	.75			.25
	SP	1.00				1.00				1.00				1.00						.75	.25	1.00			
	Crowd	.89	.11			.90	.05	.05		.88	.06	.06		1.00				.24		.65	.12	1.00			
PE_CVR_14	Faculty	1.00				.25	.25		.50	1.00				1.00						.50	.50	.75			.25
	SP	.75	.25			.75	.25			.75	.25			1.00						.75	.25	.75	.25		
	Crowd	.53	.21	.26		.75			.25	.06		.94		1.00						1.00		1.00			
PE_CVR_15	Faculty	.50	.25	.25		.75			.25	.25		.75		1.00						1.00		1.00			
	SP	.25	.50	.25		1.00						1.00		1.00						1.00		1.00			
	Crowd	1.00				1.00				1.00				1.00				1.00				1.00			
PE_CVR_16	Faculty	1.00				1.00				1.00				1.00				.75		.25		1.00			
	SP	1.00				1.00				1.00				1.00				1.00				1.00			

Table 10. Raw agreement for ICSF (item x rater type).

Item	Crowd	Faculty	SP	Item Mean
ICS_1_chx	0.96	1.00	1.00	0.99
ICS_2_chx	0.99	0.79	1.00	0.93
ICS_3_chx	0.82	0.79	0.79	0.80
ICS_4	0.33	0.38	0.46	0.39
ICS_5	0.26	0.42	0.50	0.39
ICS_6	0.27	0.58	0.42	0.42
ICS_7	0.19	0.50	0.54	0.41
ICS_8	0.37	0.63	0.54	0.51
ICS_9	0.26	0.50	0.33	0.36
ICS_10_chx	0.88	0.71	0.96	0.85
ICS_11_chx	0.78	0.67	0.88	0.77
ICS_12_chx	0.55	0.50	0.54	0.53
ICS_13	0.35	0.17	0.33	0.28
ICS_14	0.26	0.42	0.42	0.36
Chx Mean	0.83	0.74	0.86	
Global Mean	0.29	0.45	0.44	

Table 11. Average deviation for ICSF items (item x rater type).

Item	Deviation Metric	Crowd	Faculty	SP	Item Mean
ICS_4	ADabs	1.06	0.71	0.63	0.80
	ADvec	0.88	-0.29	0.21	0.27
ICS_5	ADabs	1.07	0.67	0.63	0.79
	ADvec	0.95	-0.08	0.46	0.44
ICS_6	ADabs	1.11	0.54	0.71	0.79
	ADvec	0.75	-0.38	0.38	0.25
ICS_7	ADabs	1.01	0.50	0.50	0.67
	ADvec	0.71	-0.17	0.42	0.32
ICS_8	ADabs	0.90	0.46	0.63	0.66
	ADvec	0.55	-0.38	0.46	0.21
ICS_9	ADabs	1.21	0.50	0.88	0.86
	ADvec	1.02	-0.08	0.63	0.52
ICS_13	ADabs	1.02	1.08	0.79	0.96
	ADvec	0.68	-0.25	0.38	0.27
ICS_14	ADabs	1.13	0.63	0.79	0.85
	ADvec	0.99	-0.38	0.79	0.47
Rater Mean	ADabs	1.06	0.64	0.69	
	ADvec	0.82	-0.25	0.46	

Table 12. Average deviation for ICSF items (rater type x skill level).

Rater Type	Skill Level	ADabs	ADvec
Crowd	Low	1.22*	1.09*
	High	0.76*	0.29*
SP	Low	0.70*	0.59*
	High	0.67*	0.20
Faculty	Low	0.57*	-0.18*
	High	0.77*	-0.39*

* Average deviation is significantly different from zero, one-sample t-test, $p < .05$

Table 13. Raw agreement for PESC items (item x rater type).

Item	Crowd	Faculty	SP	Item Mean
PE_1	0.68	0.75	0.88	0.77
PE_2	0.75	0.75	0.75	0.75
PE_3	0.50	0.63	0.63	0.58
PE_4	0.50	0.67	0.63	0.60
PE_5	0.74	0.92	0.92	0.86
PE_6	0.92	0.88	0.88	0.89
PE_7	0.41	0.54	0.54	0.50
PE_8	0.38	0.58	0.54	0.50
PE_9	0.54	0.63	0.54	0.57
PE_10	0.83	0.92	0.71	0.82
PE_11	0.36	0.50	0.46	0.44
PE_12	0.36	0.50	0.46	0.44
PE_13	0.37	0.46	0.38	0.40
PE_14	0.64	0.79	0.58	0.67
PE_15	0.55	0.54	0.58	0.56
PE_16	1.00	0.96	1.00	0.99
Rater Mean	0.60	0.69	0.65	

Table 14. Variance components from ICSF generalizability studies (item x rater type).

		Estimated Variance			Total Variance (%)		
Item	Variance Component	Crowd	Faculty	SP	Crowd	Faculty	SP
ICS_1_Chx	Student	0.00	*	*	0.0%	*	*
	Rater	0.00	*	*	0.0%	*	*
	Stu x Rat	0.04	*	*	100.0%	*	*
ICS_2_Chx	Student	*	0.09	*	*	47.8%	*
	Rater	*	0.01	*	*	6.0%	*
	Stu x Rat	*	0.09	*	*	46.3%	*
ICS_3_Chx	Student	0.08	0.09	0.09	47.4%	47.8%	47.8%
	Rater	0.00	0.01	0.01	0.0%	6.0%	6.0%
	Stu x Rat	0.09	0.09	0.09	52.6%	46.3%	46.3%
ICS_4	Student	0.17	0.54	0.33	20.9%	45.5%	40.7%
	Rater	0.00	0.26	0.04	0.0%	21.3%	5.4%
	Stu x Rat	0.65	0.40	0.44	79.1%	33.2%	53.9%
ICS_5	Student	0.37	1.02	0.53	32.1%	64.7%	53.8%
	Rater	0.03	0.12	0.05	2.6%	7.8%	5.0%
	Stu x Rat	0.75	0.43	0.41	65.3%	27.6%	41.2%
ICS_6	Student	0.85	1.61	1.13	51.8%	82.3%	73.5%
	Rater	0.00	0.05	0.00	0.0%	2.6%	0.0%
	Stu x Rat	0.79	0.30	0.41	48.2%	15.2%	26.5%
ICS_7	Student	0.49	0.90	0.44	34.7%	69.8%	51.6%
	Rater	0.02	0.19	0.00	1.2%	14.7%	0.0%
	Stu x Rat	0.90	0.20	0.41	64.1%	15.5%	48.4%
ICS_8	Student	1.43	1.69	2.29	63.5%	77.7%	81.5%
	Rater	0.00	0.00	0.00	0.0%	0.0%	0.0%
	Stu x Rat	0.82	0.49	0.52	36.5%	22.3%	18.5%
ICS_9	Student	0.65	1.32	0.36	43.6%	78.5%	46.9%
	Rater	0.00	0.03	0.08	0.0%	1.7%	11.0%
	Stu x Rat	0.84	0.33	0.32	56.4%	19.9%	42.1%
ICS_10_Chx	Student	0.05	0.07	0.15	36.0%	24.2%	78.3%
	Rater	0.00	0.10	0.00	0.0%	36.4%	0.0%
	Stu x Rat	0.09	0.11	0.04	64.0%	39.4%	21.7%
ICS_11_Chx	Student	0.06	0.11	0.25	37.5%	42.2%	85.7%
	Rater	0.00	0.00	0.00	1.5%	0.0%	0.0%
	Stu x Rat	0.10	0.14	0.04	60.9%	57.8%	14.3%
ICS_12_Chx	Student	0.09	0.11	0.02	39.6%	44.4%	7.8%
	Rater	0.02	0.02	0.11	7.7%	6.7%	38.8%
	Stu x Rat	0.13	0.12	0.15	52.8%	48.9%	53.4%
ICS_13	Student	0.76	0.95	0.61	45.3%	65.5%	63.6%
	Rater	0.00	0.00	0.09	0.0%	0.0%	9.9%
	Stu x Rat	0.92	0.50	0.25	54.7%	34.5%	26.5%
ICS_14	Student	0.94	1.35	0.92	47.2%	75.8%	71.0%
	Rater	0.00	0.06	0.00	0.0%	3.1%	0.0%
	Stu x Rat	1.05	0.38	0.38	52.8%	21.1%	29.0%

* Variance components not estimated due to low variability in ratings. These can be interpreted as "perfect reliability."

Table 15. Phi coefficients (single rater, max raters, cost equivalent) from decision studies for ICSF items (item x rater type).

	Crowd			Faculty			SP		
Item	Single Rater	Max Raters (k = 20)	Cost Equivalent	Single Rater	Max Raters (k = 4)	Cost Equivalent	Single Rater	Max Raters (k = 4)	Cost Equivalent
ICS_1_Chx	*	*	*	*	*	*	*	*	*
ICS_2_Chx	*	*	*	0.48	0.95	0.48	*	*	*
ICS_3_Chx	0.47	0.95	0.73	0.48	0.95	0.48	0.48	0.95	0.65
ICS_4	0.21	0.84	0.44	0.45	0.94	0.45	0.41	0.93	0.58
ICS_5	0.32	0.90	0.59	0.65	0.97	0.65	0.54	0.96	0.70
ICS_6	0.52	0.96	0.76	0.82	0.99	0.82	0.74	0.98	0.85
ICS_7	0.35	0.91	0.61	0.70	0.98	0.70	0.52	0.96	0.68
ICS_8	0.63	0.97	0.84	0.78	0.99	0.78	0.82	0.99	0.90
ICS_9	0.44	0.94	0.70	0.78	0.99	0.78	0.47	0.95	0.64
ICS_10_Chx	0.36	0.92	0.63	0.24	0.86	0.24	0.78	0.99	0.88
ICS_11_Chx	0.38	0.92	0.64	0.42	0.94	0.42	0.86	0.99	0.92
ICS_12_Chx	0.40	0.93	0.66	0.44	0.94	0.44	0.08	0.63	0.14
ICS_13	0.45	0.94	0.71	0.66	0.97	0.66	0.64	0.97	0.78
ICS_14	0.47	0.95	0.73	0.76	0.98	0.76	0.71	0.98	0.83

Note: Values are phi coefficients.
* Variance components not estimated due to low variability in ratings. These can be interpreted as "perfect reliability."

Table 16. Phi coefficients (minimum raters to reliable) from decision studies for ICSF items (item x rater type).

Item	Rater	Min. Rater to 0.7	Min. Cost to 0.7
ICS_1_Chx	Crowd	*	*
	Faculty	*	*
	SP	*	*
ICS_2_Chx	Crowd	*	*
	Faculty	3	$2.31
	SP	*	*
ICS_3_Chx	Crowd	3	$0.84
	Faculty	3	$2.31
	SP	3	$1.38
ICS_4	Crowd	9	$2.52
	Faculty	3	$2.31
	SP	4	$1.84
ICS_5	Crowd	5	$1.40
	Faculty	2	$1.54
	SP	2	$0.92
ICS_6	Crowd	3	$0.84
	Faculty	1	$0.77
	SP	1	$0.46
ICS_7	Crowd	5	$1.40
	Faculty	1	$0.77
	SP	3	$1.38
ICS_8	Crowd	2	$0.56
	Faculty	1	$0.77
	SP	1	$0.46
ICS_9	Crowd	3	$0.84
	Faculty	1	$0.77
	SP	3	$1.38
ICS_10_Chx	Crowd	5	$1.40
	Faculty	8	$6.16
	SP	1	$0.46
ICS_11_Chx	Crowd	4	$1.12
	Faculty	4	$3.08
	SP	1	$0.46
ICS_12_Chx	Crowd	4	$1.12
	Faculty	3	$2.31
	SP		
ICS_13	Crowd	3	$0.84
	Faculty	2	$1.54
	SP	2	$0.92
ICS_14	Crowd	3	$0.84
	Faculty	1	$0.77
	SP	1	$0.46

* Variance components not estimated due to low variability in ratings. These can be interpreted as "perfect reliability."

Table 17. Krippendorff's alpha coefficients for PESC items.

Item	Crowd (k = 17)	SP (k = 4)	Faculty (k = 4)
PE_1	*	*	*
PE_2	0.39	0.58	0.51
PE_3	0.69	0.47	0.59
PE_4	0.69	0.47	0.47
PE_5	0.51	0.72	0.74
PE_6	0.57	0.47	0.59
PE_7	0.22	0.26	0.42
PE_8	0.21	0.26	0.31
PE_9	*	*	0.77
PE_10	*	*	0.74
PE_11	0.52	0.31	0.35
PE_12	0.47	0.31	0.35
PE_13	0.45	0.72	0.35
PE_14	0.39	0.22	0.37
PE_15	0.67	0.76	0.46
PE_16	*	*	*

* Not possible to estimate reliability coefficient due to low variability in ratings. These can be interpreted as "perfect reliability."

Table 18. Descriptive statistics for RRX.

Item	Item Text	ICS	ICS Text Anchor	PE	PE Text Anchor
RRX_1	I felt comfortable evaluating this student's (ICS/PE) skills	5.49	Agree	5.21	Agree
RRX_2	I felt comfortable providing positive feedback about the student's social and communication skills	5.53	Strongly Agree	--	--
RRX_3	I felt comfortable providing negative feedback about the student's social and communication skills	5.29	Agree	--	--
RRX_4	I understood the elements of the questionnaire well	5.60	Strongly Agree	5.51	Strongly Agree
RRX_5	The questionnaire gave me enough information to make decisions about the student's (ICS/PE) skills	5.40	Agree	5.27	Agree
RRX_6	The instructions for completing the questionnaire were clear	5.61	Strongly Agree	5.55	Strongly Agree
RRX_7	I was motivated to make accurate ratings of this student's (ICS/PE) skills	5.57	Strongly Agree	5.88	Strongly Agree
RRX_8	I was motivated to be fair in evaluating the student's (ICS/PE) skills	5.69	Strongly Agree	5.73	Strongly Agree
RRX_9	Patients, like myself, should be able to provide feedback about the (ICS/PE) skills of future physicians	5.53	Strongly Agree	5.27	Agree
RRX_10	Patients, like myself, are capable of providing feedback about the (ICS/PE) skills of future physicians	5.40	Agree	5.14	Agree
RRX_11	I feel like I need additional training to make such evaluations in the future	2.17	Disagree	2.74	Slightly Disagree

Note: A three-color scale is used to aid interpretation of means. Full red saturation is defined at 1 (Strongly Disagree). Full yellow saturation is defined at the midpoint of the scale (3.5; Slightly Disagree/Slightly Agree). Full green saturation is defined at the 6 (Strongly Agree).

Table 19. Theme frequencies for RRX open-ended items.

Theme	ICS Task	PE Task
Task Positive	67.1%	55.8%
Task Negative	10.0%	9.1%
Technical Concerns	4.3%	1.3%
Comfortable Evaluating	12.9%	14.3%
Apprehensive Evaluating	7.1%	13.0%
Face Blurring	7.1%	0.0%
Student Performance	12.9%	14.3%
Objective	5.7%	2.6%
Enjoy/Fun	37.1%	26.0%
Interesting	20.0%	26.0%
Patient voice	28.6%	10.4%

Note: Values represent the percentage of Crowd raters making a comment belonging to a theme within task

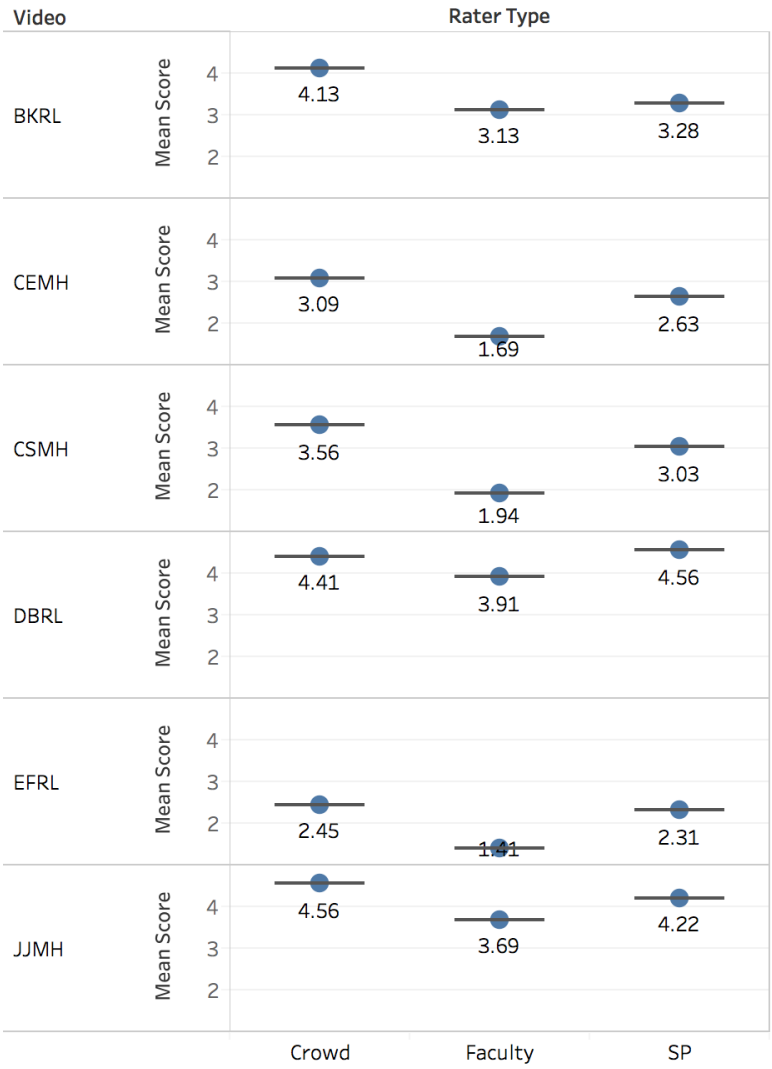


Figure 1. Mean ratings by video for ICSF global items (focus on rater type).

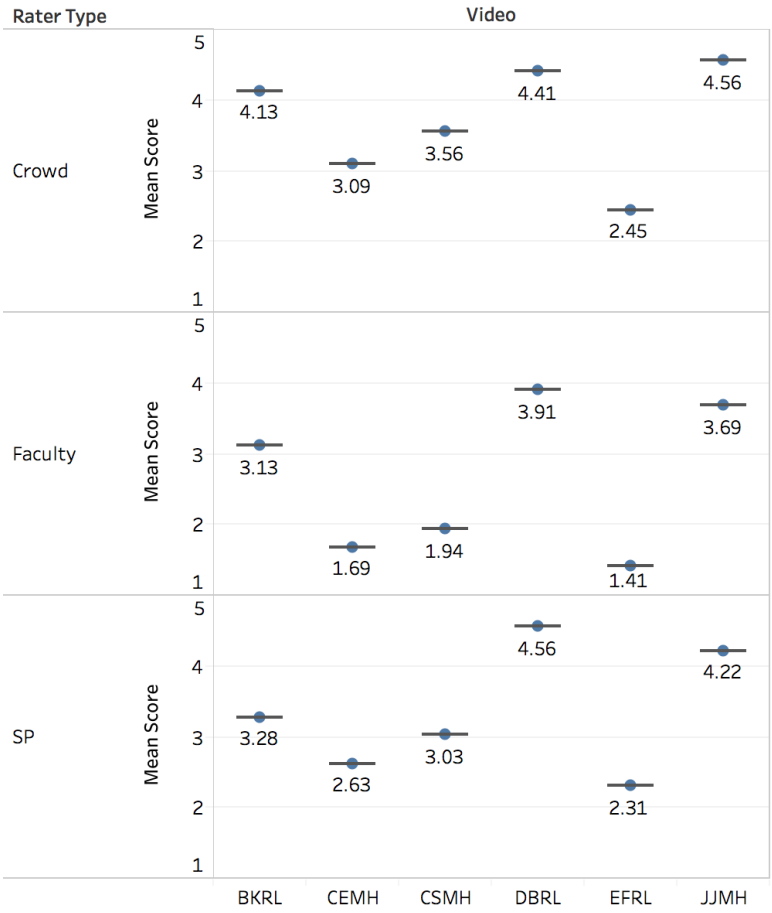


Figure 2. Mean ratings by rater type for ICSF global items (focus on video).

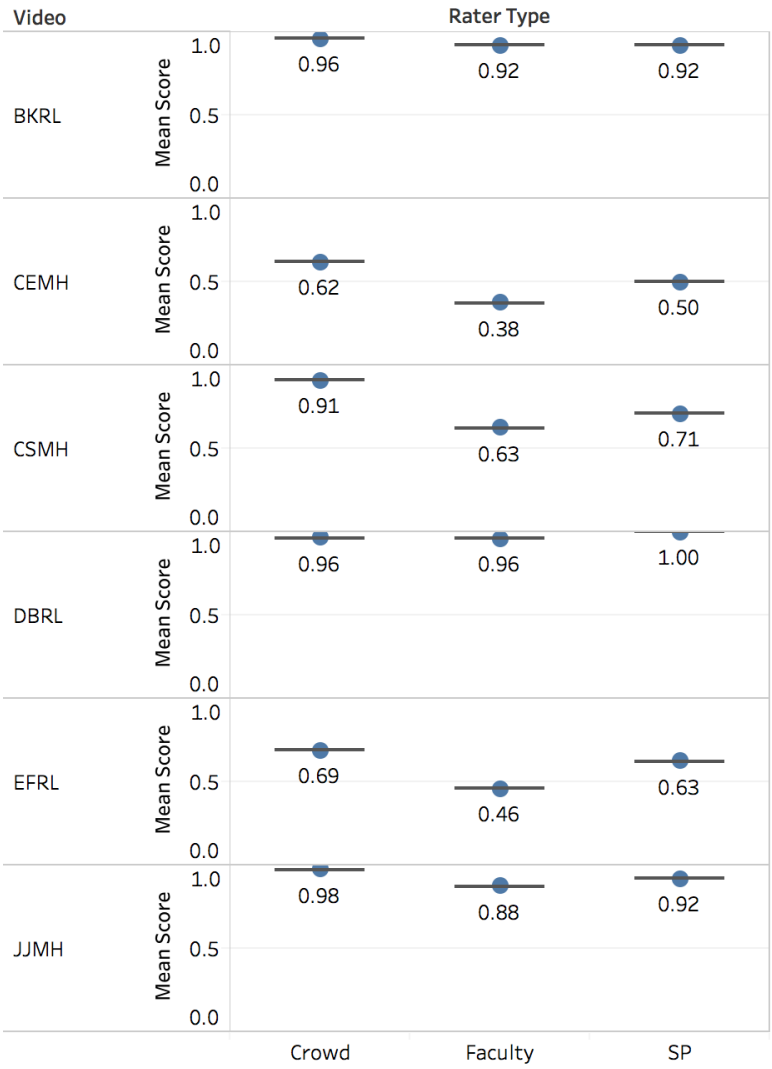


Figure 3. Mean ratings by video for ICSF checklist items (focus on rater type).

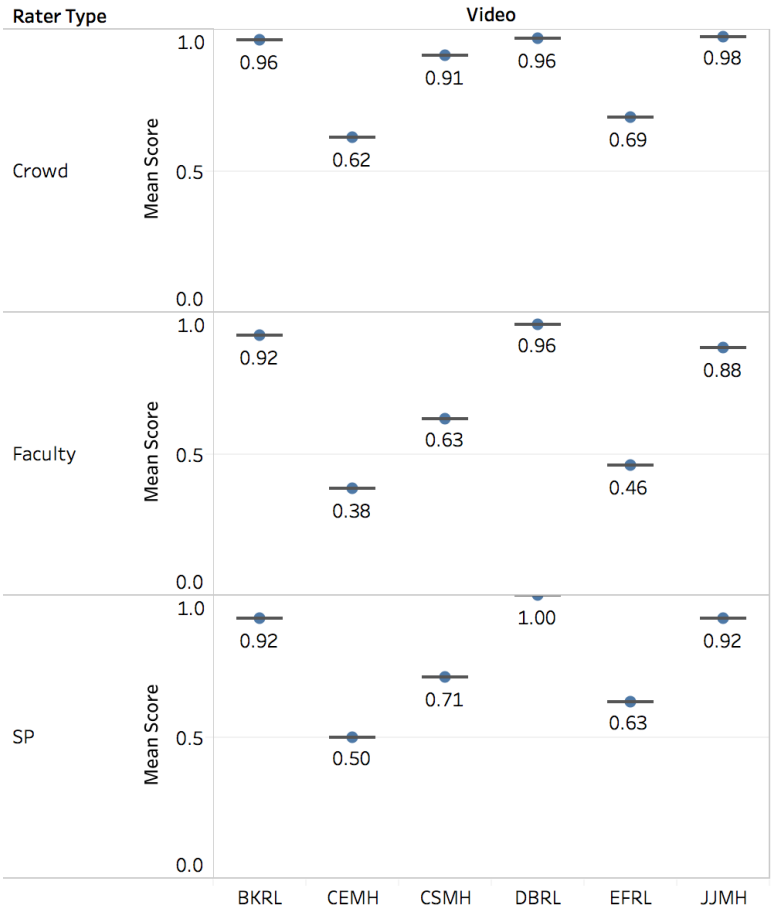


Figure 4. Mean ratings by rater type for ICSF checklist items (focus on video).

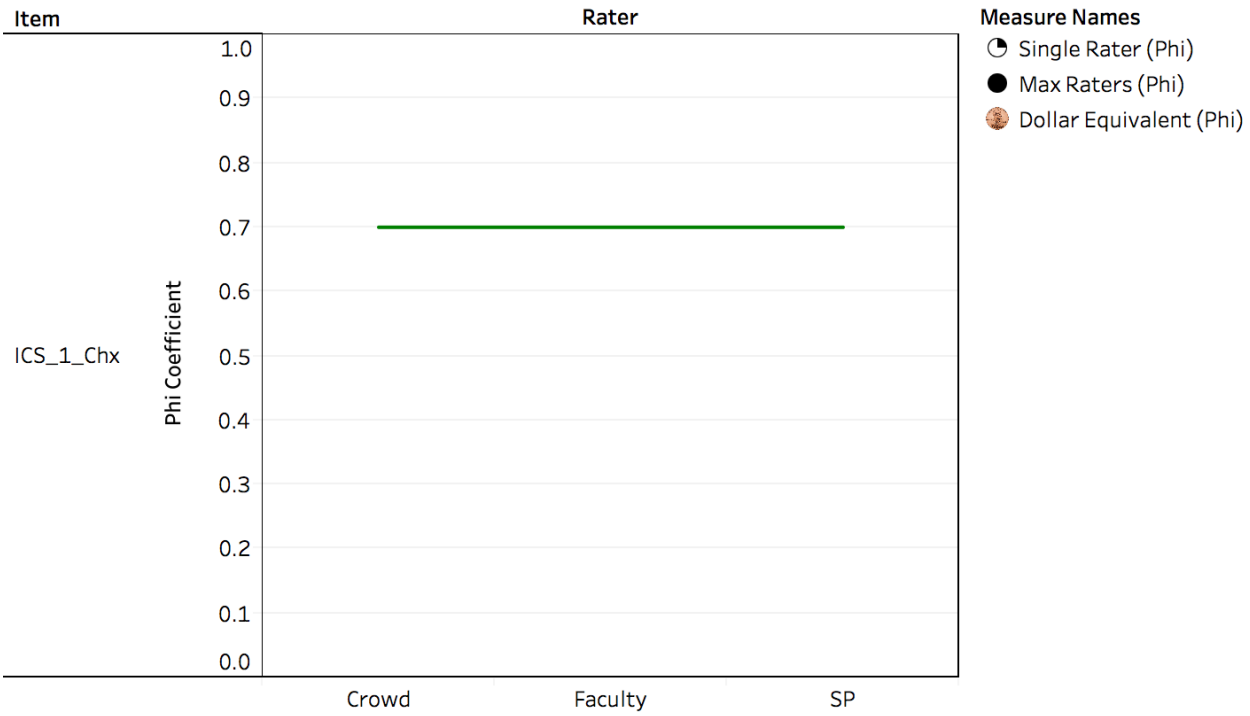


Figure 5. Decision studies for ICSF Item 1.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70. Missing data indicates an inability to conduct a G- or D-study due to low or no variability in the rating set, and should therefore be interpreted as "perfect reliability".

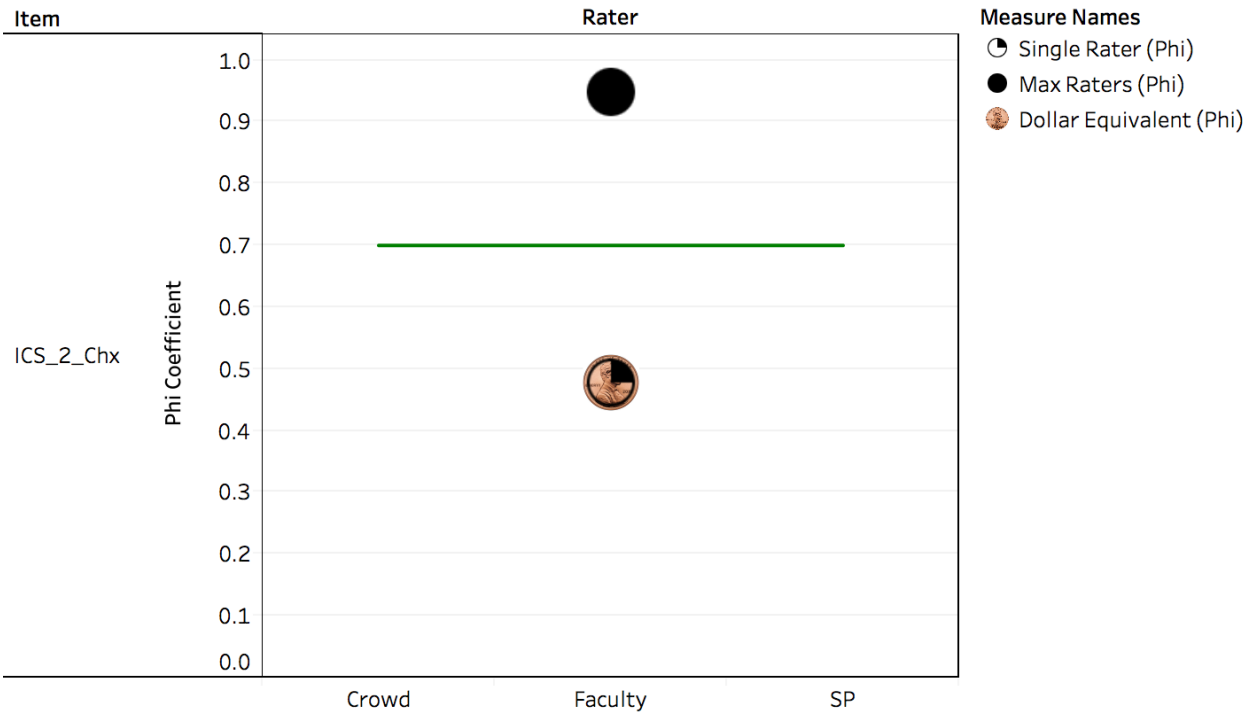


Figure 6. Decision studies for ICSF Item 2.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70. Missing data indicates an inability to conduct a G- or D-study due to low or no variability in the rating set, and should therefore be interpreted as "perfect reliability".

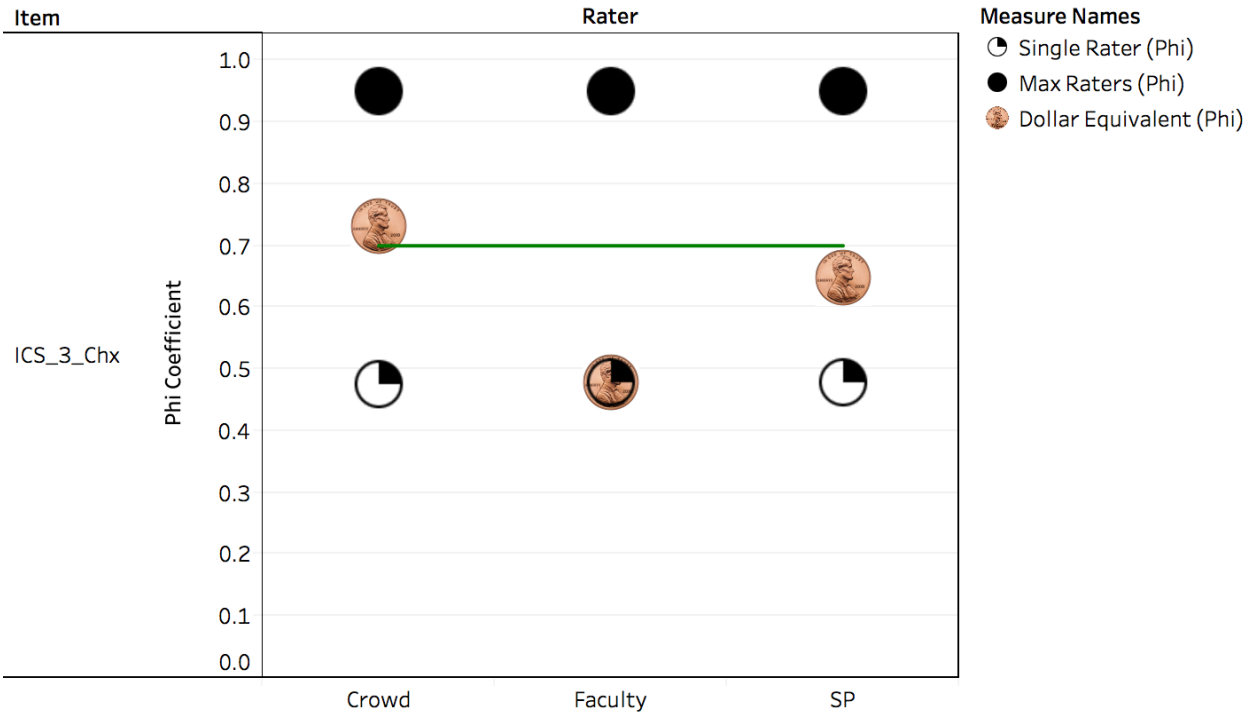


Figure 7. Decision studies for ICSF Item 3.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70. Missing data indicates an inability to conduct a G- or D-study due to low or no variability in the rating set, and should therefore be interpreted as "perfect reliability".

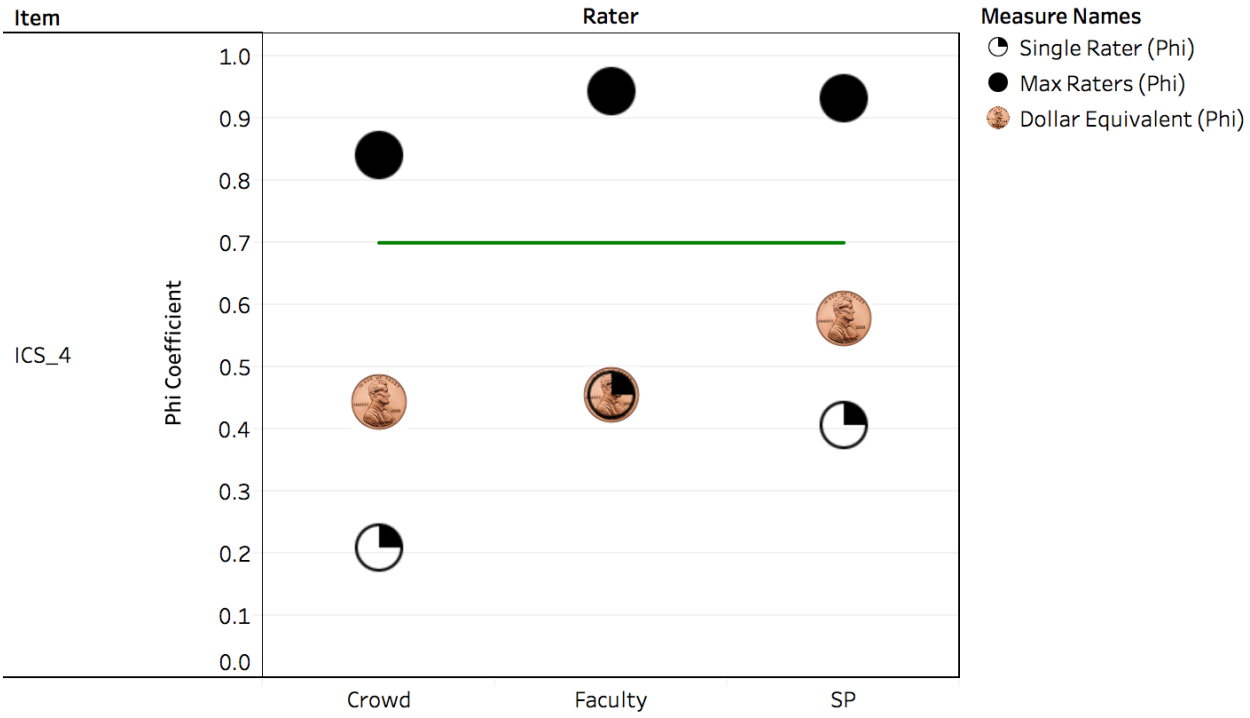


Figure 8. Decision studies for ICSF Item 4.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

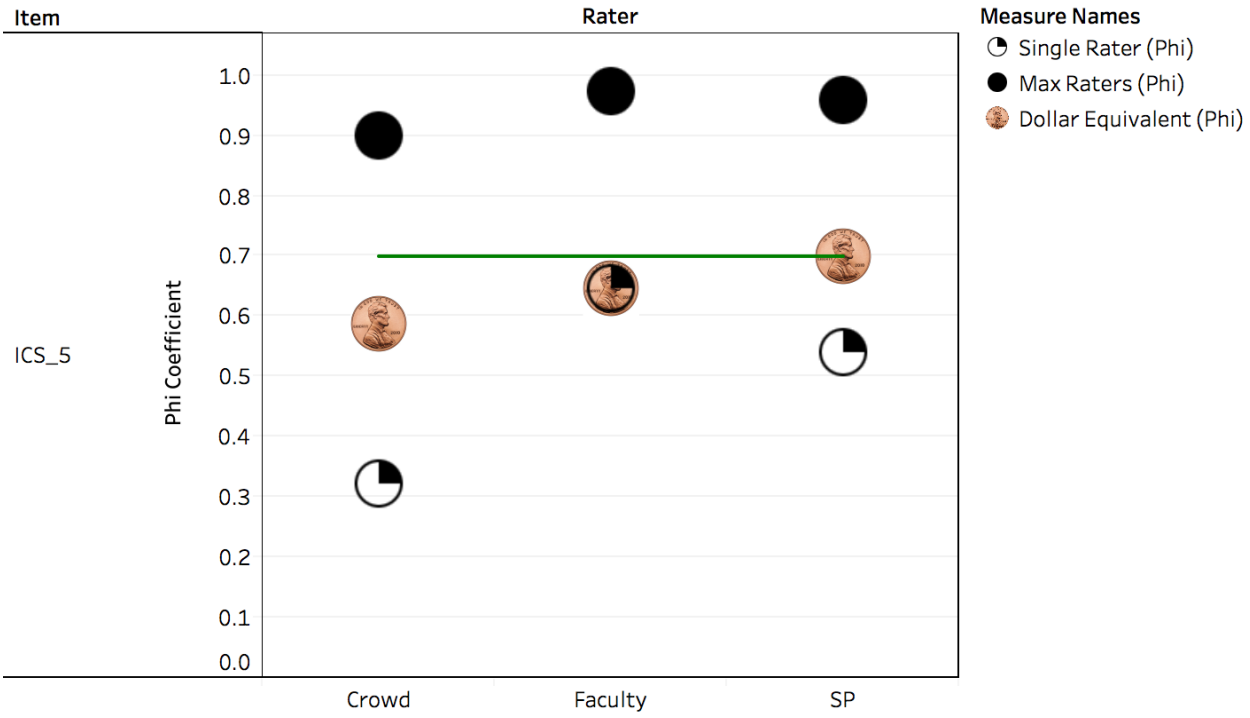


Figure 9. Decision studies for ICSF Item 5.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

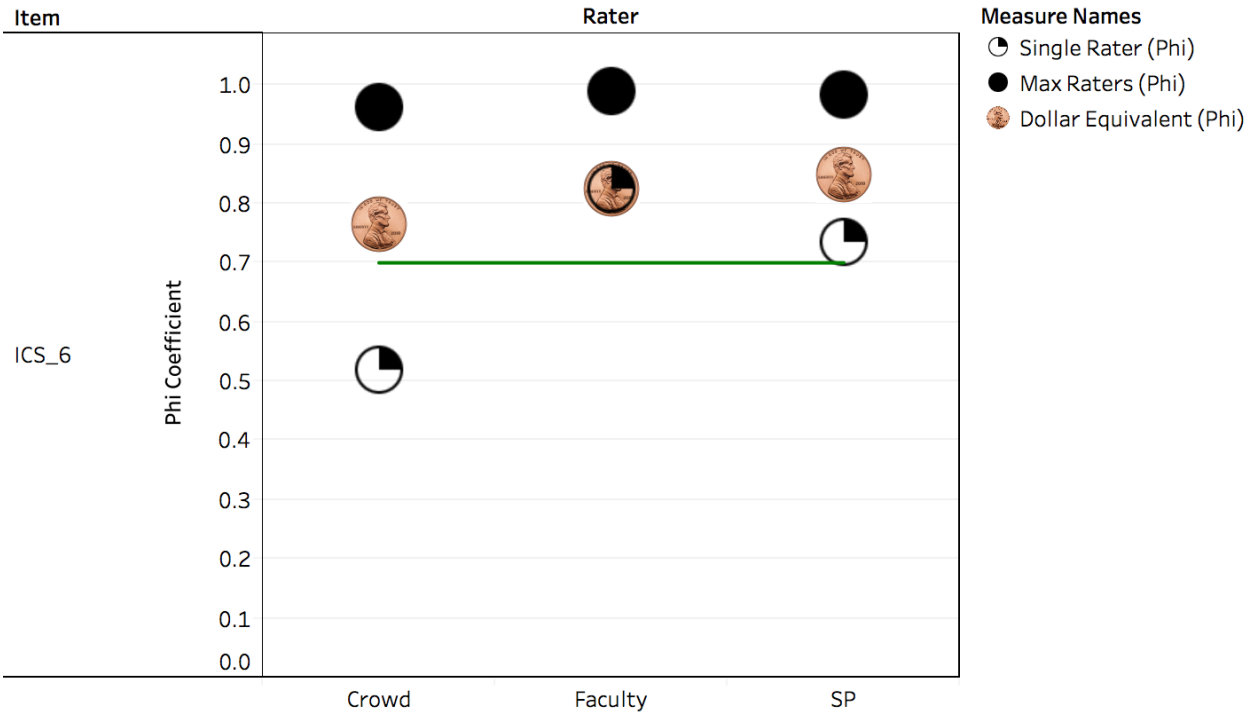


Figure 10. Decision studies for ICSF Item 6.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

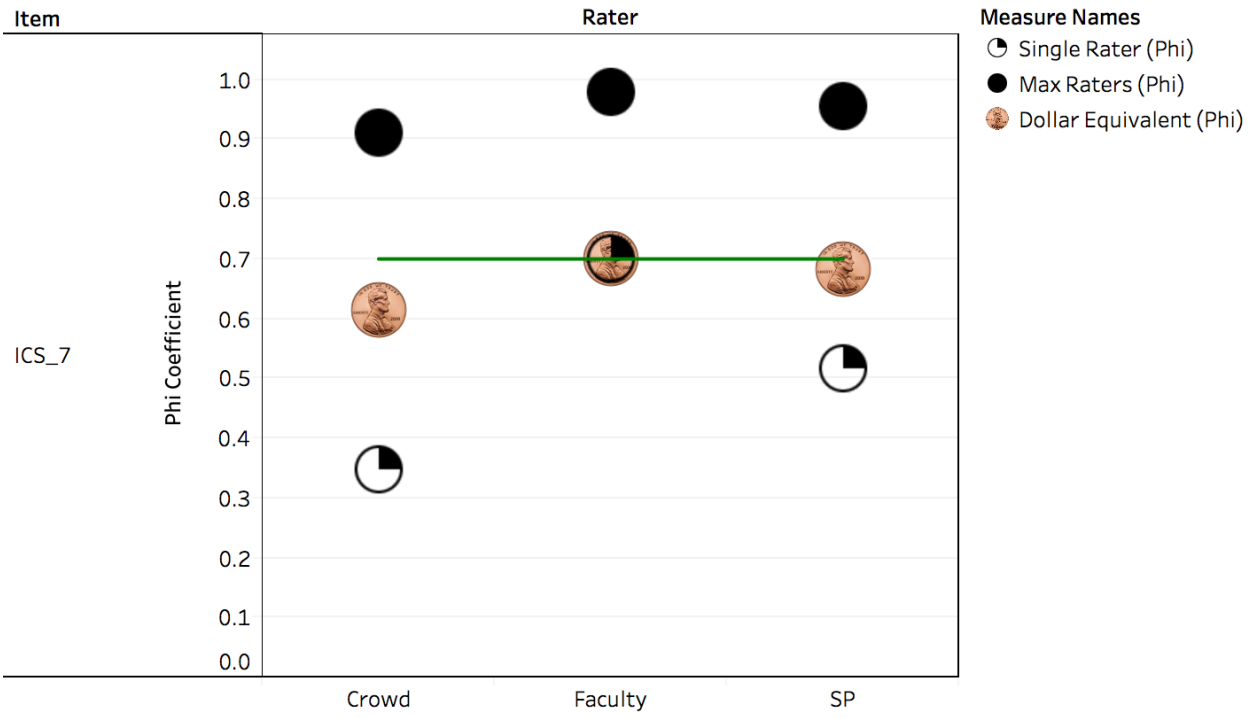


Figure 11. Decision studies for ICSF Item 7.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

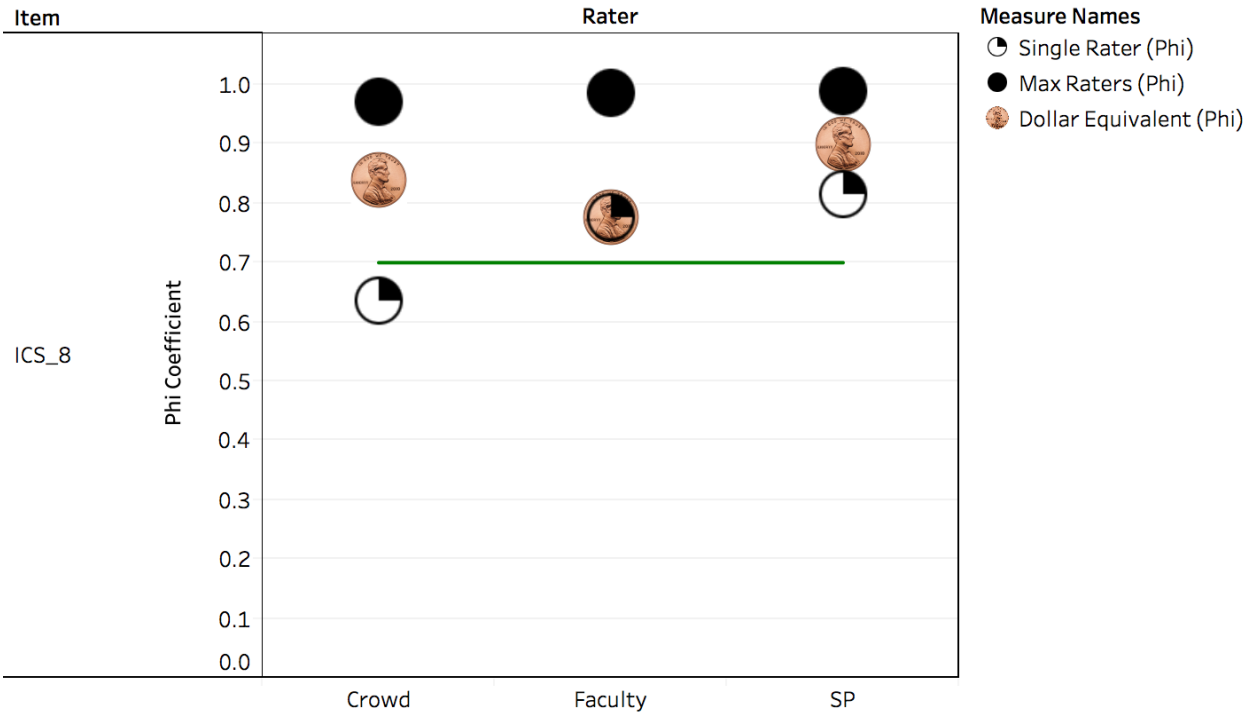


Figure 12. Decision studies for ICSF Item 8.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

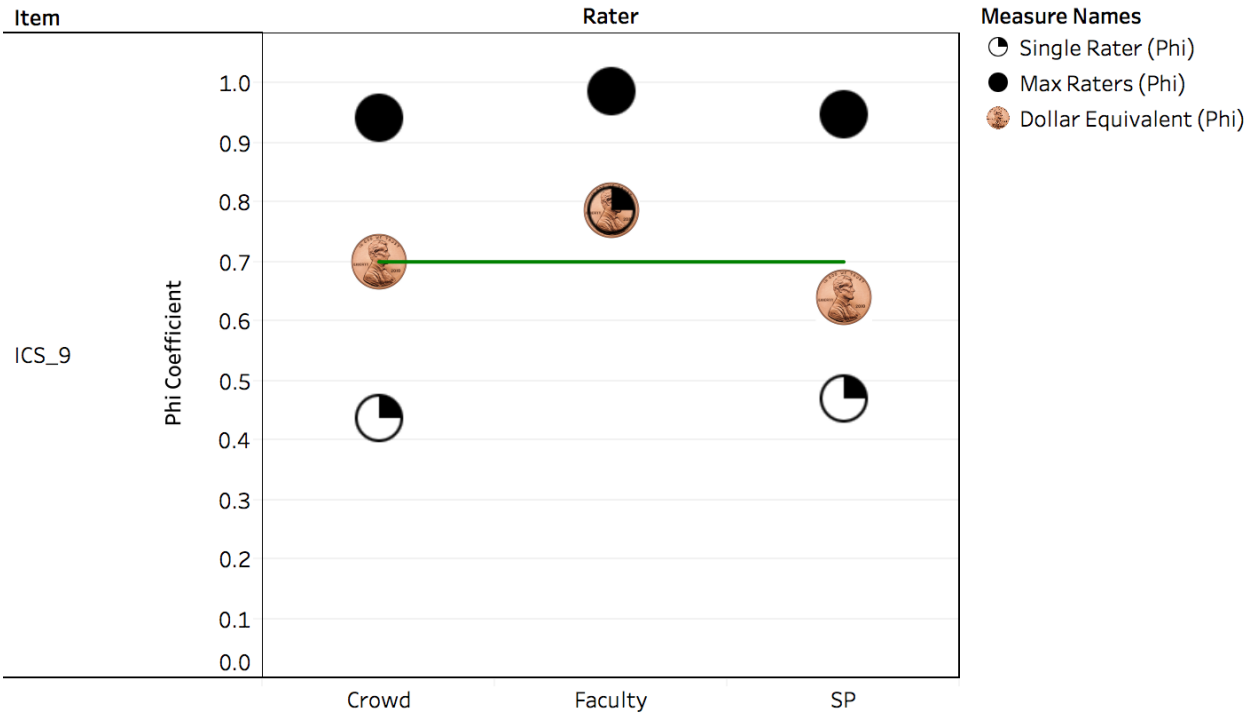


Figure 13. Decision studies for ICSF Item 9.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

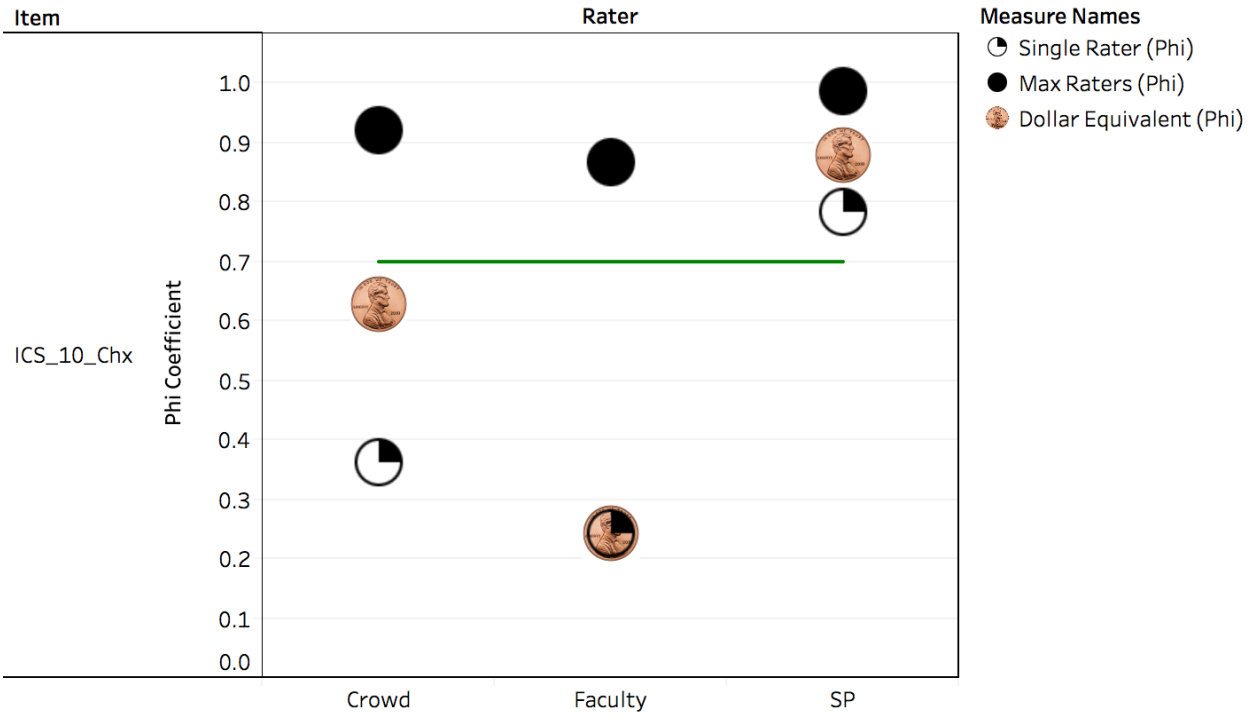


Figure 14. Decision studies for ICSF Item 10.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

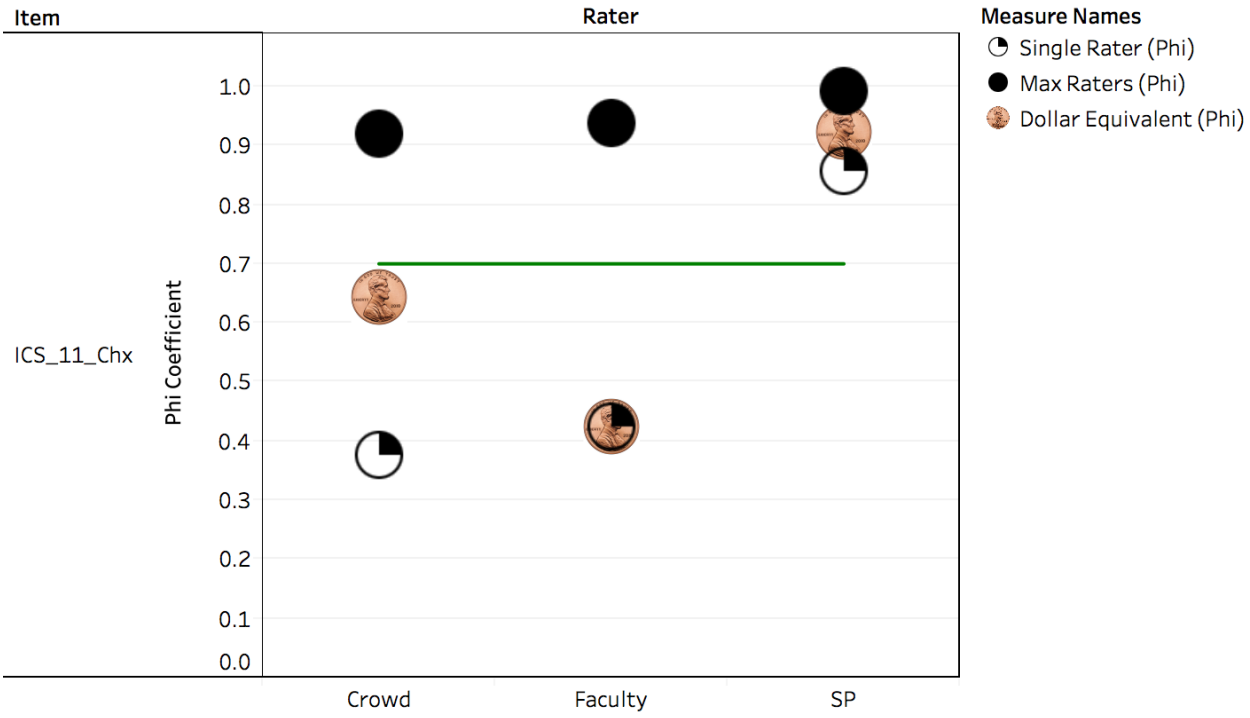


Figure 15. Decision studies for ICSF Item 11.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

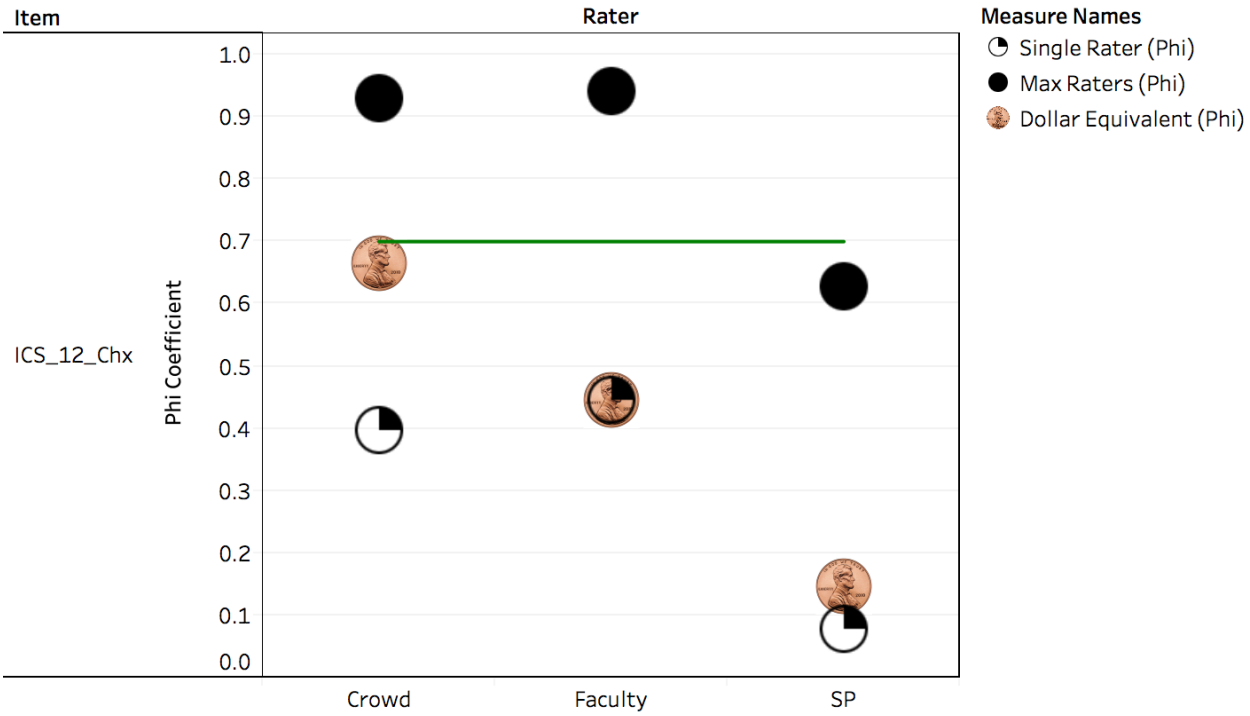


Figure 16. Decision studies for ICSF Item 12.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

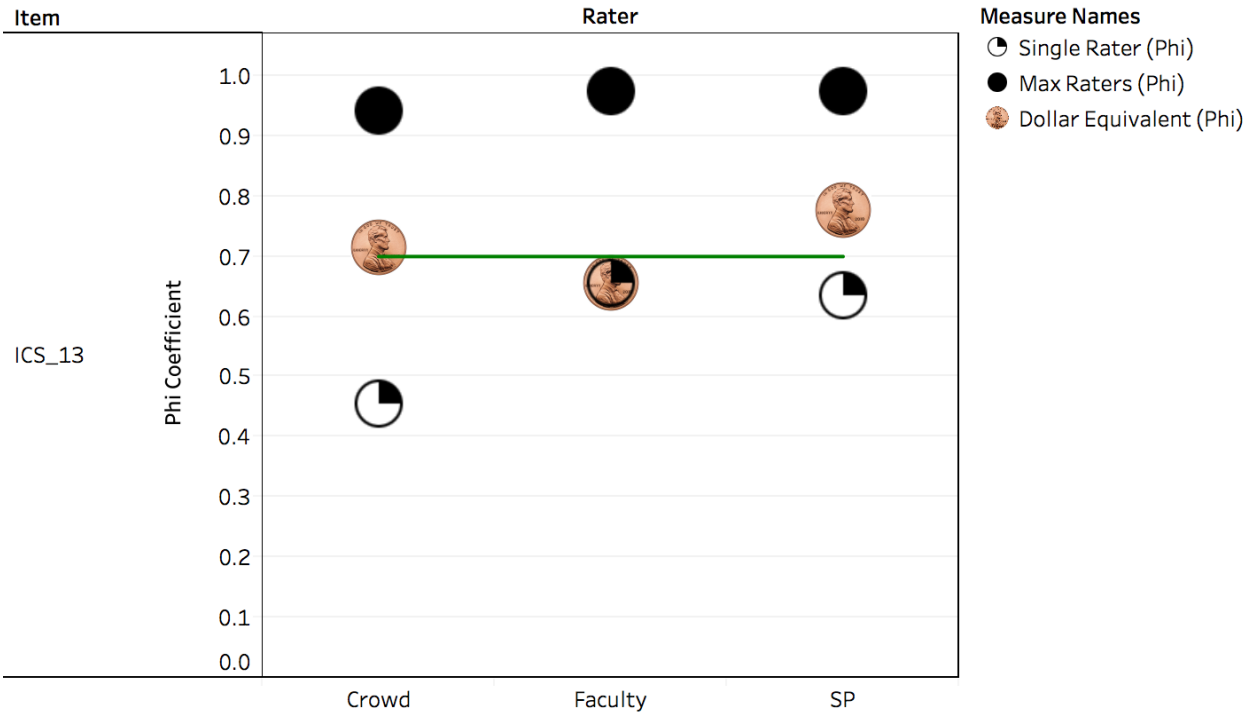


Figure 17. Decision studies for ICSF Item 13.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

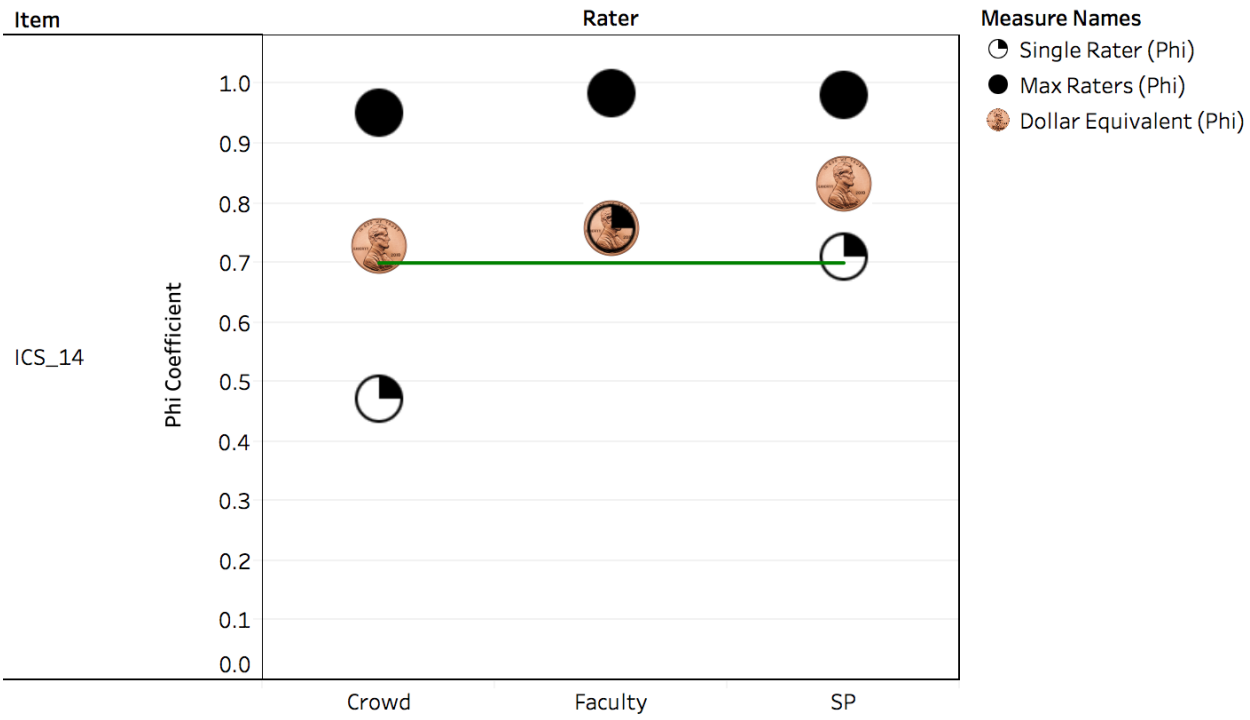


Figure 18. Decision studies for ICSF Item 14.

Note: Minimal acceptable reliability is indicated by a horizontal green bar at .70.

CHAPTER FIVE:

STUDY 2 METHOD

Participants

Study 2 participants were RMC students from all four years of the curriculum (M1-M4). All active RMC students (N = 524) were invited to participate via email. In order to ensure that perspectives of students at all levels of training were represented, I used a stratified sampling strategy. Separate tasks were created for each cohort with a maximum quota of 20 participants per cohort. A total of 81 students participated in Study 2: 21 M1's, 23 M2's, 17 M3's, and 20 M4's. Students in-progress when the quota was met were permitted to finish the survey, which explains why two of the cohorts had participation levels above quota. The average age of participants was 25.37 (SD = 2.40); the average age of all RMC students is 26.28 (SD = 2.64). Females represented 62% of the participants in this study; 50% of RMC students are female. With respect to race, 62% of student participants identified as White (57% for all RMC students), 22% as Asian (25% for all students), 5% as Black/African American (5% for all students), 4% identified with two or more races (3% for all students), and 6% reported their race as other or preferred not to answer. Students were compensated with a $15 Amazon gift card for their participation.

Measures

The student demographics questionnaire measured age, gender, and race (Appendix 6).

Two versions of the Rating and Feedback Quality Questionnaire (RFQQ) were developed to measure student perceptions of feedback package quality. The RFQQ-ICS (Appendix 17) measured the degree to which students felt that the ratings and feedback in each feedback package reflected the interpersonal and communication skills of the standardized student in the video as well as perceptions of feedback diversity, quality, specificity, and utility. The RFQQ-PE (Appendix 18) measured the degree to which students felt that the ratings in each feedback package reflected the physical exam skills of the standardized student in the video. Both forms of the RFQQ also asked the student to choose a single feedback package he or she prefers and to provide a written justification for that preference.

The Student Reactions Questionnaire (SRX; Appendix 19) was developed to measure student perceptions of present RMC OSCE feedback practices and prospective feedback practices involving crowdsourced ratings. The RRX assessed student agreement with statements about the utility, fairness, application, timeliness, privacy, motivation, and comfort with a variety of feedback practices.

Both the RFQQ and the SRX use a Likert-type scale with the following anchors: 1 = Strongly Disagree, 2 = Disagree, 3 = Slightly Disagree, 4 = Slightly Agree, 5 = Agree, 6 = Strongly Agree.

Feedback Packages

The SP, Faculty, and Crowd ratings and feedback gathered in Study 1 were used to construct feedback packages for students to evaluate in Study 2. One group of feedback packages was constructed for each task (ICS, PE) for each video. A sample of a group of feedback packages for the ICS task is presented in Appendix 20, and one for the PE task is presented in Appendix 21. Each feedback package group presented ratings from each rater type at the item

level. Feedback packages were presented side-by-side in a table with a different text color assigned to each package to facilitate student comparisons among packages. The student's own ratings and feedback of the performance in each video was presented to the left of the feedback packages.

Each feedback package came from one of the three rater types in Study 1. Feedback Package A came from SP raters, Feedback Package B came from Faculty raters, and Feedback Package C came from Crowd raters. Students were able to review the number of raters contributing ratings and feedback to each package. For the SP and Faculty raters, one of the four ratings within each rater type was randomly selected and presented in the corresponding feedback package. This random sampling procedure was repeated for each video such that each rater had an equal chance of having his/her ratings selected. Although this procedure discarded some of the ratings and feedback collected in Study 1, it was important that student reactions be examined within a framework that compares current practices at RMC (i.e., a single set of ratings from SPs and/or Faculty raters) with prospective practices (i.e., a group of ratings from Crowd raters). For feedback packages with a single set of ratings (student, SP, Faculty), the option the rater selected was presented as a single number or string. For feedback packages with multiple ratings (Crowd), a mean and standard deviation were presented for ICSF global items. For ICSF checklist and PESC items, the percentage of raters selecting each option was presented. All positive and negative narrative feedback was presented below the feedback package table, organized by rater source.

Procedure

The student task in Survey Gizmo is outlined in Appendix 22. In order ensure that reactions data was gathered from students reviewing feedback packages generated from

performance episodes across the skill continuum, a branching algorithm randomly assigned participants to providing and reviewing ratings for one of the six videos. No specific minimum or maximum quotas were defined for each video.

The first page of each rating task featured one screening question to verify that the participant was a student at RMC and in a particular cohort (M1, M2, M3, or M4). The second page featured the consent form. On the third page, students were provided with instructions and asked to review the interview and plan video and fill out the ICSF. Page four asked students to watch the physical exam video and fill out the PESC. Students watched ICS and PE videos for the same standardized student. On page five, participants reviewed the ICS feedback packages and filled out the RFQQ-ICS. On page six, students reviewed the PE feedback packages and filled out the RFQQ-PE. The corresponding videos and instruments (ICSF, PESC) were available for optional student reference at the bottom of pages five and six. Students were not informed about the source of the feedback packages when they reviewed the packages and filled out the RFQQ.

On page seven, the rater source for each of the feedback packages was revealed to the participants along with an explanation of how the ratings and feedback for each of the packages were collected. A detailed narrative highlighted the qualifications of Crowd raters and the potential advantages of a crowdsourced OSCE rating system. The benefits of treating the patient as a consumer, improved rating quality and diversity, speed, scalability, and costs were discussed. Students also had access to the entire set of instructions provided to the Crowd raters for the ICS and PE tasks. Students then filled out the SRX. Participants were only allowed to forward-navigate in the survey such that they could not revise their own video ratings or their

RFQQ responses after the feedback package sources were revealed to them. Finally, page eight collected student demographics.

Multiple attention checks were embedded in Study 2. On page three (ICS task), three items verified that the student reviewed the video: 1) a multiple-choice item about the patient's chief complaint, 2) an open-ended short answer item about the factors that aggravate the patient's condition, and 3) an open-ended short answer item about the medical student's and attending physician's recommended treatment plan. On page four (PE task), one multiple choice item asked about the type of physical exam that was being evaluated in the encounter. On pages five and six, students were asked to report the number of evaluators in each feedback package to ensure they had reviewed the ratings table. On page seven, a three-item true/false quiz verified that students read the narrative about the potential advantages of crowdsourcing OSCE ratings. Finally, a series of items embedded in the instruments (i.e., ICSF, PESC, RRX) ensured that students were not haphazardly responding (e.g., "Please select Strongly Disagree for this item").

CHAPTER SIX:

STUDY 2 RESULTS

Data Preparation

All attention checks items were inspected to identify haphazard responding. If any ratings met at least one of the following criteria, they were excluded from analyses. The number of ratings excluded based on each condition is indicated in parentheses: 1) incorrect identification of the chief complaint in the ICS task (n = 0), 2) incorrect identification of aggravating conditions in the ICS task (n = 0), 3) incorrect identification of the treatment plan in the ICS task (n = 0), 4) incorrect identification of the physical exam type in the PESC task (n = 4), 5) incorrect response to the attention check items embedded in the ICSF or PESC tools (n = 12), 6) incorrect response to more than one of three crowdsourcing explanation comprehension questions (n = 2), 7) incorrect response to the attention check items embedded in the SRX tool (n = 11), 8) misreporting the number of evaluators in each feedback package by more than two evaluators (n = 5). Several responses were disqualified based on multiple attention check criteria (n = 2). In total, 26 responses were excluded from analysis.

Rating and Feedback Quality Questionnaire (RFQQ)

Likert-type items. Each Likert-type RFQQ item was presented for each feedback package. As such, each item triplet was analyzed with a one-way repeated measures ANOVA to explore whether there were statistically significant differences in the item means for each feedback package. The independent variable, feedback package, had three levels – SP, Faculty,

Crowd. Descriptive statistics for RFQQ items are presented in Table 20. Translations from numerical scale values to narrative anchors are provided after each mean to aid interpretation. Numerical scale values were rounded to the nearest whole-number anchor.

ICS_RFQQ_1: This item measured student perceptions of the degree to which the ratings in each of the ICS task feedback packages represented performance in the ICS video. There were no statistically significant differences in agreement between the SP package (M = 4.51 (Agree), SD = 1.12), Faculty package (M = 4.22 (Slightly Agree), SD = 1.29), and Crowd package (M = 4.38 (Slightly Agree), SD = 1.24), $F(2, 108) = .97$, $p = .38$, partial $\eta^2 = .02$.

ICS_RFQQ_2: This item measured student perceptions of the degree to which the written feedback in each of the ICS task feedback packages represented performance in the ICS video. There were no statistically significant differences in agreement between the SP package (M = 4.73 (Agree), SD = 1.04), Faculty package (M = 4.55 (Agree), SD = 1.23), and Crowd package (M = 4.82 (Agree), SD = 1.11), $F(2, 108) = 1.30$, $p = .28$, partial $\eta^2 = .02$.

ICS_RFQQ_3: This item measured student perceptions of the degree to which the written feedback in each of the ICS task feedback packages was diverse. There was a statistically significant difference in agreement, $F(2, 108) = 165.73$, $p < .0005$, partial $\eta^2 = .75$. Agreement was highest for the Crowd package (M = 5.42 (Agree), SD = .71), followed by the SP (M = 2.51 (Slightly Disagree), SD = 1.30) and Faculty (M = 2.49 (Disagree), SD = 1.39) packages. Post hoc analysis with a Bonferroni correction revealed that agreement was different between the Crowd and SP packages (2.91, 95% CI [2.39, 3.43], $p < .0005$) and between the Crowd and Faculty packages (2.93, 95% CI [2.37, 3.48], $p < .0005$), but not between the SP and Faculty packages, (.02, 95% CI [-.21, .25], $p = 1.00$)

ICS_RFQQ_4: This item measured student perceptions of the degree to which the written feedback in each of the ICS task feedback packages was of high quality. There was a statistically significant difference in agreement, $F(2, 108) = 19.85$, $p < .0005$, partial $\eta^2 = .27$. Agreement was highest for the Crowd package ($M = 4.96$ (Agree), $SD = 1.20$), followed by the SP ($M = 3.96$ (Slightly Agree), $SD = 1.20$) and Faculty ($M = 3.71$ (Slightly Agree), $SD = 1.33$) packages. Post hoc analysis with a Bonferroni correction revealed that agreement was different between the Crowd and SP packages (1.00, 95% CI [.46, 1.54], $p < .0005$) and between the Crowd and Faculty packages (1.26, 95% CI [.69, 1.82], $p < .0005$), but not between the SP and Faculty packages, (.26, 95% CI [-.19, .70], $p = .49$)

ICS_RFQQ_5: This item measured student perceptions of the degree to which the written feedback in each of the ICS task feedback packages was specific. There was a statistically significant difference in agreement, $F(2, 108) = 40.51$, $p < .0005$, partial $\eta^2 = .43$. Agreement was highest for the Crowd package ($M = 5.42$ (Agree), $SD = .71$), followed by the SP ($M = 4.20$ (Slightly Agree), $SD = 1.12$) and Faculty ($M = 3.89$ (Slightly Agree), $SD = 1.15$) packages. Post hoc analysis with a Bonferroni correction revealed that agreement was different between the Crowd and SP packages (1.22, 95% CI [.76, 1.66], $p < .0005$) and between the Crowd and Faculty packages (1.53, 95% CI [1.09, 1.97], $p < .0005$), but not between the SP and Faculty packages, (.31, 95% CI [-.14, .76], $p = .28$)

ICS_RFQQ_6: This item measured student perceptions of the degree to which the written feedback in each of the ICS task feedback packages would be useful in guiding the student in the video to change his/her performance in the future. There was a statistically significant difference in agreement, $F(2, 108) = 18.03$, $p < .0005$, partial $\eta^2 = .25$. Agreement was highest for the Crowd package ($M = 5.16$ (Agree), $SD = 1.05$), followed by the SP ($M = 4.18$ (Slightly

Agree), *SD* = 1.20) and Faculty (*M* = 3.93 (Slightly Agree), *SD* = 1.27) packages. Post hoc analysis with a Bonferroni correction revealed that agreement was different between the Crowd and SP packages (.98, 95% CI [.43, 1.54], $p < .0005$) and between the Crowd and Faculty packages (1.24, 95% CI [.65, 1.82], $p < .0005$), but not between the SP and Faculty packages, (.26, 95% CI [-.21, .72], $p = .54$)

PE_RFQQ_1: This item measured student perceptions of the degree to which the ratings in each of the PE task feedback packages represented performance in the PE video. There were no statistically significant differences in agreement between the SP package (*M* = 4.38 (Slightly Agree), *SD* = 1.42), Faculty package (*M* = 4.53 (Agree), *SD* = 1.25), and Crowd package (*M* = 4.53 (Agree), *SD* = 1.10), $F(2, 108) = .35$, $p = .71$, partial $\eta^2 = .01$.

Forced choice package preference. Of the 55 students evaluating the ICS feedback packages, 18 preferred the SP package, 4 preferred the Faculty package, and 33 preferred the Crowd package. A chi-square goodness-of-fit test was conducted to determine whether the feedback packages were preferred equally. The minimum expected frequency was 18. The chi-square goodness-of-fit test indicated that student package preference was statistically significantly different ($\chi^2(2) = 22.95$, $p < .0005$), with over half of students preferring the Crowd package.

Of the 55 students evaluating the PE feedback packages, 14 preferred the SP package, 14 preferred the Faculty package, and 27 preferred the Crowd package. A chi-square goodness-of-fit test was conducted to determine whether the feedback packages were preferred equally. The minimum expected frequency was 18. The chi-square goodness-of-fit test indicated that student package preference was statistically significantly different ($\chi^2(2) = 6.15$, $p = .05$), with about half of students preferring the Crowd package.

Student Reactions Questionnaire (SRX)

Likert-type items. Several of the SRX items were naturally grouped. For example, some items had identical construction except for an inflection of a terminal word such as item SRX_17, which asked about acceptability of crowd ratings for formative assessment, and SRX_18, which asked about acceptability of crowd ratings for summative assessment. Differences in means for those items presented in pairs were analyzed using a paired-samples t-test, while those items presented in sets of three were analyzed using a one-way repeated measures ANOVA. All ungrouped items are interpreted using univariate descriptive statistics. All SRX item groupings and descriptive statistics are presented in Table 21.

SRX_1, SRX_2: These items measured whether SP or Faculty ratings would be valuable if students had ratings from the Crowd. Student agreed more that Faculty ratings would be valuable even if they had Crowd ratings (M = 5.22 (Agree), SD = .86) when compared with the value of SP ratings if students had Crowd ratings (M = 4.76 (Agree), SD = 1.20), a statistically significant mean difference of .46, 95% CI [.09, .84], $t(53) = 2.46$, p =.02.

SRX_3, SRX_4: These items measured the value of Crowd ratings if students had ratings from SPs or Faculty. There was no statistically significant difference in student agreement that Crowd ratings would be valuable if they already had SP (M = 4.91 (Agree), SD = 1.11) or Faculty ratings (M = 4.98 (Agree), SD = .95), $t(53) = .89$, p =.38.

SRX_9, SRX_10, SRX_11: These items measured student willingness to receive negative feedback from different rater types. There was a statistically significant difference in willingness, $F(2, 108) = 4.04$, p = .02, partial η^2 = .07. Willingness to receive negative feedback was highest for the Faculty (M = 5.27, (Agree), SD = .804) package, followed by the SP (M = 5.04 (Agree), SD = 1.02) and Crowd (M = 4.93 (Agree), SD = 1.10) packages. Post hoc analysis with

a Bonferroni correction revealed that willingness was different between the Faculty and Crowd packages (.35, 95% CI [.05, .64], p = .02), but not between the Faculty and SP packages (0.24, 95% CI [.00, .48], p = .05) or the SP and Crowd packages, (.11, 95% CI [-.26, .48], p = 1.00).

SRX_14, SRX_15, SRX_16: These items measured student trust in different rater types to be fair and objective in providing ratings and feedback. There were no statistically significant differences in perceptions of fairness and objectivity between the SP package (M = 4.20 (Slightly Agree), SD = 1.38), Faculty package (M = 4.56 (Agree), SD = 1.26), and Crowd package (M = 4.44 (Slightly Agree), SD = 1.24), $F(2, 108) = 2.05$, p = .13, partial η^2 = .04.

SRX_17, SRX_18: These items measured the acceptability of Crowd ratings for formative and summative assessment purposes. Students agreed that Crowd ratings would be more acceptable for formative assessment (M = 4.89 (Agree), SD = 1.10) when compared with summative assessment (M = 2.95 (Slightly Disagree), SD = 1.41), a statistically significant mean difference of 1.95, 95% CI [1.56, 2.33], $t(54) = 10.16$, $p < .0005$.

SRX_19, SRX_21: These items measured the appropriateness of Crowd raters as judges of interpersonal and communication skills and physical exam skills. Students agreed more that Crowd raters are appropriate judges of interpersonal and communication skills (M = 4.25 (Slightly Agree), SD = 1.08) when compared with physical exam skills (M = 2.89 (Slightly Disagree), SD = 1.49), a statistically significant mean difference of 1.36, 95% CI [.93, 1.80], $t(54) = 6.31$, $p < .0005$.

SRX_20, SRX_22: These items measured student comfort with Crowd rater evaluations of interpersonal and communication skills and physical exam skills. Students reported that they would feel more comfortable being evaluated by Crowd raters on their interpersonal and communication skills (M = 4.36 (Slightly Agree), SD = 1.13) when compared with their physical

exam skills (*M* = 2.85 (Slightly Disagree), *SD* = 1.52), a statistically significant mean difference of 1.51, 95% CI [1.10, 1.92], $t(54) = 7.45$, $p < .0005$.

SRX_25, SRX_26: These items measured student concerns about privacy if their faces were revealed to or obstructed from Crowd raters. Students reported that they would be more concerned about privacy issues if their face was visible (*M* = 5.20 (Agree), *SD* = .91) as compared to blurred (*M* = 3.22 (Slightly Disagree), *SD* = 1.29) in crowdsourced applications, a statistically significant mean difference of 1.98, 95% CI [1.70, 2.27], $t(54) = 13.83$, $p < .0005$.

SRX_28, SRX_29: These items measured student perceptions about SP and Faculty raters adjusting ratings in response to social forces such as avoiding negative interactions with students. There was no statistically significant difference in student agreement that Faculty (*M* = 3.15 (Slightly Disagree), *SD* = 1.21) and SP (*M* = 3.15 (Slightly Disagree), *SD* = 1.13) raters adjust their performance ratings based on social forces, $t(54) = .00$, $p = 1.00$.

SRX_30, SRX_31, SRX_32: These items measured student motivation to use ratings and feedback from different rater types to improve performance. There was a statistically significant difference in motivation, $F(2, 108) = 3.80$, $p = .03$, partial $\eta^2 = .07$. Motivation to use ratings and feedback to improve performance was highest for the Faculty (*M* = 4.98 (Agree), *SD* = 1.10) package, followed by the Crowd (*M* = 4.58 (Agree), *SD* = 1.30) and SP (*M* = 4.55 (Agree), *SD* = 1.43) packages. Post hoc analysis with a Bonferroni correction revealed that motivation was different between the Faculty and Crowd packages (.40, 95% CI [.01, .79], $p = .05$) and between the Faculty and SP packages (.436, 95% CI [.05, .83], $p = .023$), but not between the SP and Crowd packages, (.04, 95% CI [-.47, .55], $p = 1.00$).

Open-ended items. A content analysis (see methodology in Study 1) was used to analyze student comments about additional potential advantages and disadvantages of using a

crowdsourced system to review OSCE performance. The advantages and disadvantages items were analyzed separately, because different themes emerged. The percentage of cases with phrases in the category is reported in parentheses after the category name. Categories are presented in decreasing order of phrase prevalence. All comments by category for advantages are presented in Appendix 23. All comments by category for disadvantages are presented in Appendix 24.

The following categories emerged for advantages:

- Diversity/Volume (60%): Comments about the utility of receiving multiple ratings or pieces of feedback and/or the value of having diverse patient perspectives.
- Patient as Consumer (22%): Comments about the benefits of understanding the patient perspective.
- Reduction of Bias/Improvement in Quality (20%): Positive comments about accuracy, standardization, honesty, or lack of bias in crowdsourced evaluations.
- No Additional Advantages (16.4%): Comments that the student either can't think of any additional advantages or that they were all covered in the crowdsourcing narrative explanation.
- Cost (13%): Comments about the cost advantages of a crowdsourced system.
- Speed (11%): Comments about the speed advantages of a crowdsourced system.
- Specificity (7%): Positive comments about the specificity of crowdsourced feedback.
- Other (6%): Unique comments about additional potential advantages of using a crowdsourced system.

The following categories emerged for disadvantages:

- Accuracy/Quality (47%): Comments about Crowd raters not being knowledgeable about student performance requirements, potential lack of rater conscientiousness, or challenges of judging ICS performance with face obscured.
- Physical Exam Evaluation Apprehension (31%): Comments expressing apprehension about the ability of Crowd raters to evaluate physical exam skills.
- Privacy (24%): Comments expressing concerns about privacy
- Other (16%): Unique comments about additional potential disadvantages of using a crowdsourced system.
- Bias (9%): Comments expressing concern over biased raters (e.g., based on student minority status, dislike of physicians)
- No Disadvantages (6%): Comments expressing the lack of disadvantages in using a crowdsourced system
- Feedback clarification (6%): Comments expressing the lack of ability to have robust follow up discussion after receiving feedback (as one might do with a faculty evaluator)
- Feedback volume (4%): Comments expressing that the amount of feedback provided through a crowdsourced system is too voluminous.

Table 20. Descriptive statistics and mean comparisons for RFQQ.

Grouping	RFQQ Item*	RFQQ Item Text	Mean	Text Anchor**	SD
ICS_1	ICS_RFQQ_1_SP[a]	Package A: This feedback package features numerical ratings that reflect the performance you saw in the video	4.51	Agree	1.12
	ICS_RFQQ_1_Faculty[a]	Package B: This feedback package features numerical ratings that reflect the performance you saw in the video	4.22	Slightly Agree	1.29
	ICS_RFQQ_1_Crowd[a]	Package C: This feedback package features numerical ratings that reflect the performance you saw in the video	4.38	Slightly Agree	1.24
ICS_2	ICS_RFQQ_2_SP[a]	Package A: This feedback package features written feedback that reflects the performance you saw in the video	4.73	Agree	1.04
	ICS_RFQQ_2_Faculty[a]	Package B: This feedback package features written feedback that reflects the performance you saw in the video	4.55	Agree	1.23
	ICS_RFQQ_2_Crowd[a]	Package C: This feedback package features written feedback that reflects the performance you saw in the video	4.82	Agree	1.11
ICS_3	ICS_RFQQ_3_SP[a]	Package A: This feedback package features written feedback that is diverse	2.51	Slightly Disagree	1.30
	ICS_RFQQ_3_Faculty[a]	Package B: This feedback package features written feedback that is diverse	2.49	Disagree	1.39
	ICS_RFQQ_3_Crowd[b]	Package C: This feedback package features written feedback that is diverse	5.42	Agree	0.71
ICS_4	ICS_RFQQ_4_SP[a]	Package A: This feedback package features written feedback that is high quality	3.96	Slightly Agree	1.20
	ICS_RFQQ_4_Faculty[a]	Package B: This feedback package features written feedback that is high quality	3.71	Slightly Agree	1.33
	ICS_RFQQ_4_Crowd[b]	Package C: This feedback package features written feedback that is high quality	4.96	Agree	1.20
ICS_5	ICS_RFQQ_5_SP[a]	Package A: This feedback package features written feedback that is specific	4.20	Slightly Agree	1.11
	ICS_RFQQ_5_Faculty[a]	Package B: This feedback package features written feedback that is specific	3.89	Slightly Agree	1.15
	ICS_RFQQ_5_Crowd[b]	Package C: This feedback package features written feedback that is specific	5.42	Agree	0.71
ICS_6	ICS_RFQQ_6_SP[a]	Package A: This feedback package features written feedback that would be useful in guiding the student in the video to change his/her performance in the future	4.18	Slightly Agree	1.20
	ICS_RFQQ_6_Faculty[a]	Package B: This feedback package features written feedback that would be useful in guiding the student in the video to change his/her performance in the future	3.93	Slightly Agree	1.27
	ICS_RFQQ_6_Crowd[b]	Package C: This feedback package features written feedback that would be useful in guiding the student in the video to change his/her performance in the future	5.16	Agree	1.05
PE_1	PE_RFQQ_1_SP[a]	Package A: This feedback package features numerical ratings that reflect the performance you saw in the video	4.38	Slightly Agree	1.42
	PE_RFQQ_1_Faculty[a]	Package B: This feedback package features numerical ratings that reflect the performance you saw in the video	4.53	Agree	1.25
	PE_RFQQ_1_Crowd[a]	Package C: This feedback package features numerical ratings that reflect the performance you saw in the video	4.53	Agree	1.10

Note: A three-color scale is used to aid interpretation of means. Full red saturation is defined at 1 (Strongly Disagree). Full yellow saturation is defined at the midpoint of the scale (3.5; Slightly Disagree/Slightly Agree). Full green saturation is defined at the 6 (Strongly Agree).
* Significant differences between item means within a grouping are indicated by different superscript letters
** Mean is rounded to the nearest whole-number text anchor

Table 21. Descriptive statistics and mean comparisons for SRX.

Grouping	Item*	Item Text	Mean	Response Anchor**	SD
1	SRX_1[a]	SP ratings and feedback would be valuable, even if I had ratings and feedback from the crowd	4.78	Agree	1.20
	SRX_2[b]	Faculty ratings and feedback would be valuable, even if I had ratings and feedback from the crowd	5.22	Agree	0.86
2	SRX_3[a]	Crowd ratings would be valuable, even if I had ratings and feedback from SPs	4.91	Agree	1.11
	SRX_4[a]	Crowd ratings would be valuable, even if I had ratings and feedback from faculty	4.98	Agree	0.95
	SRX_5	Crowdsourced ratings and feedback are worthwhile	4.83	Agree	1.11
	SRX_6	Generally, I receive a sufficient amount of ratings and feedback about my clinical skills as part of simulated patient encounters	3.71	Slightly Agree	1.33
	SRX_7	The ratings and feedback I generally receive as part of simulated patient encounters give me the information I need to help me improve my performance	3.67	Slightly Agree	1.38
	SRX_8	The ratings and feedback I receive from SPs as part of simulated patient encounters are typically an accurate representation of my performance	3.75	Slightly Agree	1.19
3	SRX_9[a]	I am willing to receive negative feedback from SPs, even if the comments might upset me	5.04	Agree	1.02
	SRX_10[b]	I am willing to receive negative feedback from faculty, even if the comments might upset me	5.27	Agree	0.80
	SRX_11[a,b]	I am willing to receive negative feedback from crowd raters, even if the comments might upset me	4.93	Agree	1.10
4	SRX_12	It is important that I get feedback from multiple sources about my clinical skills	5.15	Agree	0.96
	SRX_13	The patient's perspective about my clinical skills is important	5.36	Agree	0.91
	SRX_14[a]	I trust that SP raters are fair and objective when providing ratings and feedback	4.20	Slightly Agree	1.38
	SRX_15[a]	I trust that faculty raters are fair and objective when providing ratings and feedback	4.56	Agree	1.26
	SRX_16[a]	I trust that crowd raters would be fair and objective when providing ratings and feedback	4.44	Slightly Agree	1.24
5	SRX_17[a]	Crowdsourced ratings and feedback would be acceptable for formative assessment purposes (i.e., for practice/feedback)	4.89	Agree	1.10
	SRX_18[b]	Crowdsourced ratings and feedback would be acceptable for summative assessment purposes (i.e., scores count toward final grade/promotions decisions)	2.95	Slightly Disagree	1.41
6	SRX_19[a]	I feel that crowd raters are appropriate judges of interpersonal and communication skills	4.25	Slightly Agree	1.08
	SRX_21[b]	I feel that crowd raters are appropriate judges of physical examination skills	2.89	Slightly Disagree	1.49
7	SRX_20[a]	I would feel comfortable being evaluated by a crowd rater on my interpersonal and communication skills	4.36	Slightly Agree	1.13
	SRX_22[b]	I would feel comfortable being evaluated by a crowd rater on my physical examination skills	2.85	Slightly Disagree	1.52
	SRX_23	It is important for me to receive ratings and feedback in a timely manner	5.16	Agree	0.86
	SRX_24	Generally, I receive ratings and feedback from my simulated patient encounters in a timely manner	3.51	Slightly Agree	1.32
8	SRX_25[a]	I would be concerned about privacy issues if my simulated clinical encounter performance video was evaluated by crowd raters with my face visible	5.20	Agree	0.91
	SRX_26[b]	I would be concerned about privacy issues if my simulated clinical encounter performance video was evaluated by crowd raters with my face blurred	3.22	Slightly Disagree	1.29

Note: A three-color scale is used to aid interpretation of means. Full red saturation is defined at 1 (Strongly Disagree). Full yellow saturation is defined at the midpoint of the scale (3.5; Slightly Disagree/Slightly Agree). Full green saturation is defined at the 6 (Strongly Agree).
* Significant differences between item means within a grouping are indicated by different superscript letters
** Mean is rounded to the nearest whole-number text anchor

Table 21 (Continued)

Grouping	Item*	Item Text	Mean	Response Anchor**	SD
	SRX_27	The loss of information (e.g., about eye contact) a crowd rater experiences when my face is blurred is a worthy tradeoff to protect my privacy	4.75	Agree	1.42
9	SRX_28[a]	I believe SP raters sometimes adjust performance ratings and feedback up or down due to social forces (e.g., avoiding negative interactions with students)	3.15	Slightly Disagree	1.13
	SRX_29[a]	I believe faculty raters sometimes adjust performance ratings and feedback up or down due to social forces (e.g., avoiding negative interactions with students)	3.15	Slightly Disagree	1.21
10	SRX_30[a]	I am motivated to use the SP ratings and feedback to improve my performance	4.55	Agree	1.42
	SRX_31[b]	I am motivated to use the faculty ratings and feedback to improve my performance	4.98	Agree	1.10
	SRX_32[a]	I would be motivated to use the crowdsourced ratings and feedback to improve my performance	4.58	Agree	1.30

CHAPTER SEVEN:

DISCUSSION

The aim of this dissertation was to evaluate a crowdsourced system for gathering OSCE ratings and feedback. Several potential advantages of a crowdsourced system were proposed, including the ability to decrease rating costs while maintaining rating quality, to provide learners an increased volume and diversity of ratings, to deliver feedback to students in a timelier manner, and to consider the benefits of including the patient perspective as students develop their clinical skills. Two studies were conducted to answer a series of research questions. The first study explored the costs and time associated with collecting Crowd ratings for two types of tasks. The first task was evaluating interpersonal and communication skills using an instrument predominantly composed of global rating scale items as well as several checklist items, the ICSF. The second task was evaluating physical examination skills using a technical checklist, the PESC for the cardiovascular exam subcomponent. The first study compared the accuracy and reliability of Crowd ratings with the current state-of-the-art, ratings from SPs and Faculty. Study 1 also measured Crowd rater reactions in order to understand how confident and efficacious raters felt performing the two evaluation tasks. The second study presented the ratings and feedback gathered in the first study to medical students. In Study 2, medical students performed the two evaluation tasks themselves, assessed the quality of the ratings and feedback gathered from the Crowd, SP, and Faculty raters, and provided their reactions about current OSCE rating and feedback practices as well as the prospects of a crowdsourced OSCE system.

Who is "the Crowd"?

Crowd raters were representative of the US population with respect to race, but were younger, more educated, more female, and earned less than the typical individual in the US population. Considering that median per-person healthcare spending for those 65 and older is nearly five times that of median per-person healthcare expenditures for those younger than 65 (Agency for Healthcare Research and Quality, 2014), Crowd patients are likely not representative of the US patient population. These demographic results are consistent with previous findings (Sheehan & Pittman, 2016). It may be tempting to buy into the mythology of tropes like the grouchy and cantankerous old curmudgeon or the sweet and kind grandmother, and therefore limit the generalization of Crowd patient evaluations to those types of patients represented in the Crowd rater pool. However, future research should investigate whether these demographic differences (e.g., age, gender, income) are in fact associated with systematic differences in their expectations of or preferences for physician behavior.

How Long Did It Take and How Much Did It Cost to Collect Crowd Ratings?

The best estimate of how long it would take a rater to complete the rating task was represented by Total Task Time, calculated as the total amount of time raters spent on the survey page with the video and instrument. Provided the same videos and instruments, Mean Total ICS Task Time for SP and Crowd raters was substantially higher than for Faculty raters. Crowd raters task times were likely longest because they also had to read instructions on the same page, familiarize themselves with the rating instrument, and because they were inexperienced with the task. The difference in SP and Faculty task times is likely due to the fact that RMC Faculty regularly provide ICSF ratings through video review, whereas SPs at RMC provide ICSF ratings from memory after case portrayal. Thus, Faculty and SP raters were likely faster than Crowd

raters because they are familiar with the ICSF and have developed robust mental models of student behaviors that map to ICSF domains. In conversations with some Faculty raters, I understand that Faculty members often fill out the ICSF *while* they watch the video. This anecdotal evidence may be another explanation why Faculty raters were faster than SP raters. For the PE task, the difference in Total Task Time between SP and Faculty raters was negligible, whereas Crowd raters took approximately three times as long as the other two rater types. This is unsurprising, because Crowd raters had a long set of instructions to read through and evaluating a highly technical physical exam is considerably less commonplace for lay people than evaluating social skills.

The most striking timing data was the time it took to get all of the rating packages completed. The Crowd packages were completed within hours. However, it took 1-2 weeks to receive ratings back from a far fewer number of SP and Faculty raters. These timing results are consistent with the surgical crowdsourcing studies (e.g., Holst et al., 2015). Thus, crowdsourcing OSCE ratings could realistically allow students to receive a large volume of feedback within a short period of time. It is important to note that this study only put a limited amount of "load" on the MTurk system (i.e., 6 HITS with 20 assignments each). It would be important to study timing using a load more characteristic of an operational system (for RMC, that might be approximately 140 HITs – one HIT/student). Due to resource constraints, such as the number of available exam rooms and SP actors to portray the case, simulation events typically take place over a few days. In order to distribute load on the crowdsourcing system, it would be useful to investigate alternate strategies such as posting all of the student videos for each day as the event progresses. It is also important to note that this was a voluntary research study. Therefore, the time it took

Faculty and SPs to return ratings may not be representative of those in an operational environment.

With respect to compensation, the total MTurk costs in this study are not representative of the fees that might be paid to raters in an operational environment, because the study asked MTurk workers to complete elements beyond the rating task (e.g., consent, RRX, demographics). There was substantial variability in the Total Task Time for Crowd raters, with the standard deviation of some videos near 50% of mean Total Task Time for that video. This within-video variability in timing can likely be attributed to an array of factors such as the number of times raters reviewed portions of the video, rater effort and pace, and review strategy (e.g., watch video then provide ratings, provide ratings while watching video). Compensation is considered a property of the HIT and is determined prior to posting the assignment. Since compensation is linked to task completion rather than time spent, it is in a worker's best interest to finish the assignment as fast as possible in order to increase the effective pay rate. MTurk requesters are encouraged to provide time estimates when posting, but slower raters are paid the same amount as faster raters. Given the variability in task timing, I expected that those raters who spent more time than advertised in the HIT would be dissatisfied. However, the only Crowd rater comments that mentioned compensation, praised the task for paying well.

It is essential to further study the central tendency of task completion time for Crowd raters, particularly as repeat raters begin to bring down task completion times because they are familiar with instructions. Since it would be logistically complicated to create separate tasks for new vs. repeat raters, it might be advisable to create a qualification task that presents instructions and a practice rating task. Then only those raters who have participated in the qualification task can participate in actual rating tasks. This way, the requester is not overcompensating repeat

raters by paying them for the time they built into the task for potential new raters to review instructions.

Were Crowd Ratings Accurate and Reliable?

Over ninety percent of the Crowd ratings collected met the most basic criteria expected of any evaluation, that the evaluator pay attention to the performance episode and instrument. This was achieved using relatively loose MTurk screening criteria, a minimum of 500 HITs completed with a 97% approval rate. Research ethics dictated that even those Crowd raters who did not pass attention checks still be compensated. However, an applied Crowd rating system would not have such a restriction and could be engineered to be relatively waste-free. That is, Crowd raters who do not clear attention checks would not be compensated. MTurk also allows requesters to create qualifications that MTurk workers can earn. For example, if a requester were to post a HIT that involved transcribing audio from an interview, the requester could set a condition that only workers who had earned a typing speed qualification through a separate typing speed task could work on the transcribing task. Similarly, future Crowd rating tasks could require workers to earn an OSCE rating task qualification by, for example, asking workers to rate a short series of videos with known levels of performance and only qualifying those workers with acceptable levels of accuracy. Asking MTurk workers to earn qualifications costs more, but those costs may be offset through increased rating quality (thus necessitating fewer ratings) and decreased administrative costs associated with time spent rejecting submitted assignments.

Rating accuracy was assessed against gold-standard ratings provided by the RMC Manager of Simulation Education, an SP by training. The picture of overall ICSF rating accuracy was complex. With respect to ICSF raw agreement at the item level, no single rater type consistently agreed more with true score ratings, although Crowd raters seemed to show up less

frequently in lists of highest raw agreement. The level of raw agreement also appeared to be highly dependent on the specific item being evaluated, ranging from a low of 28% for a global item to a high of 99% for a checklist item. When aggregating across items, Faculty raters agreed with true scores less than other rater types for checklist items, but no differences across rater types were found when separating raw agreement for ICSF checklist items by the skill level of the student in the video. Crowd raters agreed less with true scores than other rater types for global items; this effect appears to be driven by differences in the skill level of the student in the video as Crowd raters had significantly lower proportions of raw agreement with global rating true scores when evaluating low skill students. RA for checklist items was higher than for global items, which is expected because the global items had more performance levels available (5) to choose from than the checklist items (2).

Rating accuracy for the ICSF global items was also examined using Average Deviation metrics. With respect to absolute deviation, Crowd raters deviated more from true scores than SP and Faculty raters. A vectored average deviation metric was used to characterize the direction of the deviation and revealed that Crowd raters were on average just under one full performance level (on a scale of 1-5) more lenient relative to true score, SP raters were about half a performance level more lenient relative to true score, and Faculty raters were about a quarter of a performance level harsher relative to true score. Analysis within each rater type by skill level of the student in the video revealed that Faculty rater accuracy was consistent regardless of the level of the student's skill whereas both Crowd and SP raters were less accurate when viewing low skill (as compared to high skill) videos.

For PESC items, Faculty raters had the highest number of items (8 of 16) with the highest raw agreement. The remaining items had either equal rates of raw agreement across several rater

types or had the highest raw agreement when rated by SP or Crowd raters. Raw agreement rates across individual items varied widely (.40-.99). When analyzed by skill level of the student in the video, raw agreement proportions across raters were similar. However, raw agreement rates for low skill videos were lower for Crowd raters than for Faculty raters.

With respect to single-rater reliability for the ICSF items, Faculty members tended to have the highest levels of reliability for global items and SP raters tended to have the highest levels of reliability for the checklist items. Single Crowd-raters yielded few items with acceptable levels of reliability. However, when the maximum number of raters available in this study were included, almost all items for all rater types reached high levels of reliability. After controlling for cost, SP raters tended to be the most reliable dollar-for-dollar, although more than half of global items met acceptable levels of reliability for all rater types after controlling for cost and about half of checklist items met acceptable levels of reliability for Crowd and SP raters after controlling for cost. Finally, to reach acceptable levels of reliability, it was least expensive to use SPs for the majority of the global items whereas half of the checklist items would have been cheapest with Crowd raters and half with SP raters. The PESC reliability results were mixed with optimal rater choice dependent on the particular item.

Taken together, the accuracy results indicate that Faculty were generally more accurate and reliable than SP and Crowd raters. Crowd raters typically provided the least accurate ratings when compared with the other rater types, but the absolute levels of accuracy for Crowd raters were reasonable and in many cases not too different than the levels exhibited by other rater types, particularly when examining the lowest and highest levels of accuracy across items with rater types (i.e., SP and Faculty raters weren't perfect either).

It is also interesting to see that Crowd raters were lenient relative to internal raters, and that Faculty raters may be our students' harshest critics. Perhaps a relevant analogy is a baseball player training with a doughnut. The baseball doughnut is a weighted ring that players attach to their bats during warmup. The weighted ring functions as a desirable difficulty in training. When removed during actual gameplay, the player's muscles, used to activation levels with a weighted bat, allow for faster bat velocity when using an unweighted bat. We may be similarly overtraining students in interpersonal and communication skills by using more stringent raters during simulated patient encounters. This may lead to levels of patient satisfaction and compliance (e.g., in sharing diagnostically-relevant information, adhering to treatment plans) beyond those predicted by internal assessments. Given that the global items were anchored using detailed behavioral descriptions, the rating differences between rater types is noteworthy and may indicate that despite anchors and instructions, raters use the rating instruments in different ways.

With respect to reliability, Crowd raters lag behind internal rater types when holding other conditions constant (i.e., single rater, cost equivalence). However, it is important to note that groups of Crowd raters are capable of reaching high levels of reliability with additional ratings. Furthermore, although SP raters had higher levels of reliability under cost equivalence conditions, many of the Crowd-rated items had acceptable levels of reliability when holding costs equal. A visual examination of the figures comparing rater types under cost equivalence conditions also shows that Crowd raters are not far behind other rater types and that adding a nominal amount (e.g., 1-2) of additional Crowd raters would allow Crowd raters to reach reliability levels that would equal or exceed those of SP or Faculty raters. Although this would cost more than employing internal raters, it is important to keep these tradeoffs in mind as I

discuss other benefits of using crowdsourced ratings. That is, the additional cost of recruiting more Crowd raters to reach levels of reliability that equal fewer SP or Faculty ratings might be justified provided the additional benefits that can only be achieved by using a crowdsourced system.

The accuracy and reliability ranges across items are of particular interest. Unlike many examinations of OSCE performance presented in the literature, this study examined rating performance at the item level citing that item-level analyses reflect the common application of the ICSF and PESC instruments in the educational environment. Specifically, while there is a role for the use of these instruments at the scale level for the purposes of rank-ordering students and making promotion decisions, the more common and impactful use of these instruments is to provide students with detailed feedback they can use to improve their performance. The differences in accuracy and reliability found in this study justify future analysis at the item level, because it is clear that all global items and all checklist items are not rated uniformly. Educators and administrators should also consider targeting interventions to improve the reliability of ratings at the item level. For example, those items involving stethoscopes had particularly low accuracy and reliability. This may be due to difficulty in visually discriminating between the bell and diaphragm of the stethoscope in a video-recorded performance episode. Rating accuracy and reliability could perhaps be improved if the bell and stethoscope were associated with different colors (e.g., wrapping the bell using colored tape) rather than asking raters to rely on size cues.

How Did Crowd Raters Feel About the Rating Tasks?

Across both the ICS and PE tasks, Crowd raters agreed that they felt comfortable evaluating the student's skills, that the questionnaire gave them enough information to make decision about the student's skills, and that patients are capable of providing feedback about the

ICS and PE skills of future physicians. Crowd raters strongly agreed that they understood the elements of the questionnaire well, that the instructions for completing the questionnaire were clear, that they were motivated to make accurate ratings of the student's performance, and that that they were motivated to be fair in making their ratings. Raters agreed that patients should be able to provide feedback about PE skills and strongly agreed that they should be able to provide feedback about ICS skills. Raters slightly disagreed that they would need additional training to make PE ratings in the future, and disagreed that they would need additional training to assign ICS ratings. For the ICS task, raters strongly agreed that they were comfortable leaving positive written feedback and agreed that they were comfortable leaving negative written feedback. These results suggest that raters generally felt quite efficacious and comfortable evaluating student clinical skills, although they felt slightly more confident in the ICS task.

A set of follow-up analyses divided RRX responses according to the skill level of the student in the ICS or PE video they watched. Specifically, there was a trend for raters to react more poorly (although still positively) when evaluating low skill videos. For example, raters felt less comfortable evaluating low skill PE videos than high skill videos, raters felt more confident that they had enough information to make decisions about a student's performance when evaluating high skill PE and ICS videos compared to low skills ones, raters perceived instructions to be less clear in low skill PE videos, raters viewing low skill PE videos felt less confident that patients should be able to, and are capable of, providing feedback about PE skills, and raters viewing low skill videos felt that they needed more training for future evaluations. Interestingly, these rater perceptions are consistent with the accuracy results presented above and suggest that, as a group, raters have some insight into their limitations and performance capabilities. If Crowd ratings are used as a primary method to evaluate students, faculty and

administrators would not know the performance level of the student in each video prior to submitting them for evaluation using the crowdsourced system. If accuracy and rater feelings of efficacy are driven by student skill level, this is cause for concern because evaluation quality is likely to matter more for lower-performing students who need to adjust a larger repertoire of their behaviors. It is not immediately clear which strategies might be used to resolve this differential.

When reviewing ICS videos, raters viewing low skill videos reported higher levels of motivation to be accurate and fair. Although these results did not replicate for the PE task, these findings suggest that Crowd raters take their roles seriously. Low skill videos are likely more challenging to rate because they deviate from model behavior and require the rater to make more nuanced decisions about student performance. Thus, raters seem to "rise to the challenge" such that the more challenging the rating task is, the more they may dedicate themselves to doing a good job.

Reviewing Crowd rater comments revealed that a majority of raters found the task and instructions to be clear and straightforward. PE raters made comments about their apprehension to evaluate at about twice the rate of ICS raters; this is consistent with the numerical ratings found in the RRX. However, about 13% of raters in each task made comments indicating they were comfortable evaluating. 37% of ICS raters and 26% of PE raters left comments that indicated they enjoyed the task or had fun, while 20% of ICS and 26% of PE raters commented that the task was interesting. The largest gap in the percentage of comments was for patient voice. Approximately 29% of ICS raters commented about the virtue of patients participating in physician training (e.g., "Working on this task was a very satisfactory personal experience for me. I was really happy that I could help future doctors.") whereas only about 10% of PE raters

expressed the same sentiment; this differential is unsurprising in light of the RRX numerical ratings. In summary, rater comments generally expressed positive Crowd rater sentiment about evaluating medical student performance, with a larger proportion of positive comments for the ICS task. The comments are encouraging for the prospects of building a real crowdsourced OSCE rating system, and several raters even directly expressed a desire for more tasks like this to be made available on MTurk (e.g., "Interesting hit, I enjoyed the task and hope more will be available in the future.")

Did Students Prefer Feedback from Crowd Raters?

After students reviewed all of the feedback packages, but before they knew the source of each feedback package, students slightly agreed or agreed that the numerical ratings in the ICS feedback packages from each rater source represented the performance they saw in the video, although none of these differences were significantly different. Students also agreed that the numerical ratings in all of the PE packages represented the performance they saw in the video. Students agreed that the written feedback in all of the ICS feedback packages represented the performance in the video. Students perceived the written feedback in the ICS Crowd package to be more diverse, more specific, of higher quality, and more useful in guiding performance improvement than the written feedback in the SP and Crowd packages. When asked to choose a single ICS package, 60% of students chose the Crowd package, 33% chose the SP package, and 7% chose the Faculty package. When asked to choose a single PE package, 50% chose the Crowd package and 25% each chose the SP and Faculty packages.

Although these results are promising, it is important to remember that the packages were presented to students using existing RMC conditions (i.e., one rating from an SP or Faculty rater) and the Crowd ratings were presented under idealized conditions (i.e., the maximum number of

Crowd ratings collected in this study). As such, the Crowd packages presented to students would cost more to collect than the SP and Faculty packages. As most of the comments justifying the selection of the Crowd package tended to be related to the volume of ratings, these positive findings may not necessarily be the result of students valuing raters from Crowd raters, per se, but of a general desire to receive more feedback, regardless of source. Furthermore, when asked to think of advantages for a crowdsourced system, 60% of students commented about the desirability of a higher volume and/or increased diversity in ratings. Despite the differences in rater accuracy and reliability in the numerical ratings that were found in Study 1, students perceived all of the feedback packages to be equally representative of the performance in the videos. Again, this may not be an even comparison because the analyses in Study 1 examined reliability holding all other variables (e.g., number of raters, cost) equal whereas Study 2 compared single-rater SP and Faculty ratings (condition which showed moderate reliability in Study 1) with max-rater Crowd ratings (a condition which showed high levels of reliability). If volume, regardless of rater source, is the most important element for students, Crowd raters still represent the most inexpensive source of ratings.

How Did Students React to Crowd Ratings?

After revealing the source of each feedback package and explaining the rationale that drove the collection of crowdsourced ratings, students filled out the SRX. With respect to general feedback preferences and current RMC practices, students agreed that it is important for them to get feedback from multiple sources about their skills, but only slightly agreed that they currently receive a sufficient amount of feedback, that the feedback they receive gives them enough information to improve their performance, and that feedback from SPs accurately represents their performance in simulated patient encounters. They agreed that it is important they get ratings and

feedback returned in a timely manner but only slightly agreed that they get them back in such a manner at RMC. In open-ended comments about the advantages of a crowdsourced system 13% of students noted potential cost benefits and 11% noted potential speed benefits.

With respect to Crowd ratings specifically, students agreed that the patient's perspective about their clinical skills is important, that Crowd ratings are worthwhile, and that they would value receiving Crowd ratings even if they already had feedback from SPs and Faculty. 22% of students wrote open-ended comments about the advantages of understanding the patient's perspective as a consumer of the student's behavior. However, students also agreed that SP and Faculty ratings, in particular, would be important if they already had Crowd ratings. Students agreed that Crowd ratings would be appropriate for formative purposes, but only slightly agreed that they would be acceptable for summative use.

Students slightly disagreed that Crowd raters are appropriate judges of PE skills and that they would feel comfortable being evaluated by Crowd raters on their PE skills, but slightly agreed that Crowd raters would be appropriate for and that they would feel comfortable with ICS evaluations. 31% of students made comments expressing their apprehension about Crowd raters evaluating PE skills. It is difficult to corroborate their concern in light of the inconclusive PE accuracy and reliability results. Students were presented with the instructions Crowd raters saw for the PE task, but they were presented as an optional link that students could click on. Since this was toward the end of the student task and there was evidence of rater fatigue (12 student responses were eliminated because they missed simple within-instrument attention checks on the same page), it is unclear whether students actually saw the amount of explicit PE instruction provided to Crowd raters.

When comparing among rater types, students slightly disagreed that SP and Faculty raters adjust their ratings due to social forces; they agreed that Faculty raters are fair and objective when assigning ratings but only slightly agreed that SP and Crowd raters are or would be fair and objective, although these differences were not statistically significant. 20% of students made comments about the advantages of a crowdsourced system in mitigating bias. However, 10% of students also made comments about crowdsourced ratings as a potential source of bias.

The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council of Measurement in Education, 2014) specify that it is incumbent on the instrument designer to provide clear guidelines to assessment users about the appropriate uses of the assessment, including the provision of a clear rationale for the assessment, support for the intended use, and guidance about anticipated misuse of the assessment. It was clear from several student comments that some parts of the narrative package that attempted to explain crowdsourcing provided by the assessment developer (myself) were not clearly understood by the assessment user (the students). For example, a comment such as, "Non medical people do not understand the line of questioning or how questions [should] be phrased" betrays the sentiment implied in the ICSF about information gathering and sharing that the patient, as a consumer, should be offered insight about the line of questioning and that their emotional experience of how questions are phrased matters. As another example, a comment such as, "Moreover, the people who participate in the crowdsourcing may hold biases," does not take into account that portion of the narrative that explained the benefits of aggregating multiple ratings to disperse the potential impacts of bias in single ratings. Perhaps the narrative was ineffective in explaining this psychometric principle, or perhaps there is a misguided student belief that internal raters are free from bias. There is a clear

need for using a rich qualitative approach to develop an understanding of the motivations and boundary conditions of student attitudes about crowdsourcing. This type of inquiry can in turn lead to construction of robust student and faculty development programs that provide support for and caution against the use of crowdsourced assessments of clinical skill using rationales and pieces of evidence that are meaningful to those users.

Students slightly disagreed that they would be concerned about privacy if their faces were blurred. However, students agreed that they would be concerned about privacy if their faces were visible in Crowd rater evaluations and that the loss of ICS information from face blurring would be a worthy tradeoff to protect their privacy. In comments about potential disadvantages, 24% of students expressed concerns about privacy, including several comments about obscuring their voice.

In addition to dispersing the effects of bias through the aggregation of ratings, technological interventions such as video blurring or voice modulation offer an opportunity to mitigate the impact of biases driven by student individual differences such as gender, race, and attractiveness. In addition to respecting student preferences for privacy, educators should carefully consider whether or not to use these technology-driven tweaks. For certain educational objectives, educators and students may want bias-free estimates of student performance. However, for those objectives that concern the preparation of students for genuine clinical environments, educators may want to preserve ecological validity, because real patients have real biases.

Students agreed that they are willing to receive negative feedback and are motivated to apply the feedback they receive from all rater types, but especially from Faculty. Interestingly, in open-ended comments about the potential disadvantages of a crowdsourced system, 47% of

students expressed their concerns about the potential limitations of accuracy or quality of crowdsourced ratings. While these concerns are consistent with the actual accuracy and reliability studies in Study 1 for single-rater arrangements, they are unfounded when using ratings from intermediate-size groups of Crowd raters. They are also inconsistent with the fact that the majority of students preferred the Crowd package before they knew the source of each of the feedback packages! The student concerns about rater conscientiousness seem to be inconsistent with the effort dedicated by the majority of Crowd raters, as evidenced by the low number of ratings that had to be removed from analysis, RRX ratings indicating that raters were motivated to be accurate and fair, and Crowd rater comments about striving to do a good job in light of their understanding of the import of the task.

Although not necessarily representative of student reactions at large due to low frequency, I would like to highlight several student comments that bring up interesting ideas about crowdsourced OSCE ratings that were not considered in the formulation of this dissertation.

The first comment, "if every school uses [a] crowdsourced system, these reviews may have national standardized potential" takes the idea of crowdsourced ratings beyond application at a single institution. The only OSCE-style assessment administered at the national level is USMLE Step 2 Clinical Skills (CS). However, students typically sit for that examination towards the end of their undergraduate medical training, which means that the time left for a school to offer additional training resources is quite restricted. Furthermore, since Step 2 CS is intended to be a licensing exam, the performance reporting is broad. If a group of schools committed to a common set of cases and instruments to administer during some of their OSCEs, they could share a common pool of Crowd raters (e.g., MTurk). This would allow for the establishment of

cross-institutional metrics that would allow both individual students and programs to benchmark themselves.

Another interesting comment read, "…in terms of privacy, I would even be a little concerned about people recognizing my voice... especially if this became such a common thing that people knew that you could look on this website to see all medical students interviewing SPs." This introduced privacy concerns when a crowdsourced system is deployed at scale. Although the voice concern is technologically simple to deal with, the comment brings up a challenge of controlling the rater pool. For example, some of the MTurk raters happened to be medical students or residents. If RMC students knew that RMC consistently posted student performance episodes on MTurk, there is no mechanism available on MTurk to prevent RMC students from rating videos of other RMC students. Although the de-identification procedures used in this study may be adequate to protect student privacy from the public in general, they may not be enough to prevent peers from recognizing one another. However, the crowdsourcing literature offers a potential solution to this problem, a managed crowd. Managed crowdsourcing solutions employ a curated set of crowd workers. For example, a managed crowd for OSCE ratings might exclude students from the target institution employ only patients, and allow for the curation of a crowd that represent the demographics of the patient population the institution serves. Although curated crowds tend to be more expensive, they allow crowd workers to develop cumulative experience in a particular task and requesters to apply quality metrics and deliver feedback to crowd workers on an individual basis (Ben Christensen, 2015).

Limitations and Future Directions

This study examined the viability of a crowdsourced rating system within the ecosystem of single institution. Even the RMC students sampled in this study are likely to have had

heterogeneous OSCE feedback experiences. These differences are likely driven both by planned changes over time in the simulation program or curriculum across student cohorts as well as through individual differences such as individual student performance and traits (e.g., feedback receptivity, learning orientation) interacting with other elements in the ecosystem (e.g., raters). The variability in SRX item responses (e.g., "Generally, I receive ratings and feedback from my simulated patient encounters in a timely manner) is presented as evidence of this heterogeneity across students.

As such, the need for and success of a crowdsourced rating system is likely to be ultimately dependent on the particular components already present in and expected of the OSCE rating and feedback systems in individual institutions. Specifically, I suspect that variables such as the specific learning activities and measurement tools, the quantity, quality, sources (e.g., Faculty, SPs, residents), and formats (e.g., written narratives, guided oral debriefs) of feedback as well as the institutional norms, operational costs, and opportunities for honest feedback from patients will move the needle in favor of or against crowdsourced rating systems.

Faculty and SP cost calculations did not take into account the considerable amount of resources required to train and calibrate these internal raters, so the true cost of SP and Faculty ratings are likely higher than reported in this manuscript. On the other hand, SP ratings can be less expensive than reported here, depending on the rating model used by an institution. For example, SPs are often asked to perform two tasks simultaneously during an encounter, portray the case and assess student performance. Such active observation arrangements are financially and logistically attractive, because they appear to obviate the need for scheduling and paying for additional SP time and maintaining facilities for offline video review to complete student assessments. However, serving double duty comes at a cost as active observation increases SP

cognitive load and decreases accuracy, when compared with passive observation (i.e., video review only; Newlin-Canzone et al., 2013).

Although the SRX was rather comprehensive, concerns about rater fatigue precluded the separation of a large portion of SRX items into finer-grained inquiries about attitudes towards PE or ICS feedback. For example, students may have quite different attitudes about PE and ICS feedback with respect to, "It is important that I get feedback from multiple sources about my clinical skills" or "I trust that faculty raters are fair and objective when providing ratings and feedback." In consideration of the stark differences in student affect for those items where PE and ICS *were* separated, further inquiry is clearly warranted, because a detailed understanding of student feedback needs and reservations is at the heart of evaluating the prospects of a crowdsourced system.

As several comments from Crowd raters and students underscored, facial affect is an important "channel" in interpersonal interactions. Given student reactions in Study 2, a crowdsourced system for ICS is more likely to be acceptable to students than one for PE, and a system using face blurring is more likely to be acceptable than one that leaves faces visible. However, providing accurate estimates of performance in those domains likely to be influenced by diction *as well as* facial affect (e.g., empathy) is important as students will communicate with their faces unobscured in real life patient encounters. Therefore, it is essential to systematically study any potential effects of face blurring on rater perceptions of student performance across ICS domains. It would also be interesting to see how little ICS data is necessary to maintain the validity of ICS ratings. For example, if one presented only audio-recorded performance episodes, which ICS domains would still be reasonably assessed and which would suffer?

Although this study explored student reactions to crowdsourced feedback, it is also essential to explore faculty reactions to feedback from Crowd raters. This manuscript began by exploring the notion of expertise. The apprenticeship tradition in medical education is prevalent, and students acknowledge faculty as masters. This is clearly evidenced by the SRX data, as faculty judgment about performance is valued above all other feedback sources by medical students. Therefore, buy-in from faculty members is likely a key ingredient for the success of crowdsourced ratings. For example, if faculty members discount the value of crowdsourced ratings during coaching or feedback sessions, this type of social modeling may lead students to similarly devalue the feedback and not incorporate it into their learning plans. On the other hand, if faculty teach students effective strategies for mining the higher volume, more diverse, and less polished crowd feedback for meaningful themes, students would be more likely to incorporate crowd feedback into their performance management plans.

Another potentially fruitful avenue to explore is the relationship between student individual differences and crowdsourced feedback receptivity. For example, learning goal orientation (e.g., performance vs. mastery) has been shown to moderate the relationship between feedback receptivity and feedback characteristics such as specificity and valence (Waples, 2015). Similarly, it is possible that individual factors such as openness to experience or learning goal orientation can moderate the relationship between reactions to crowdsourced feedback and features such as how polished, constructive, or actionable it is. For example, students with a mastery-orientation may be accepting of crowd feedback regardless of tone, whereas performance-oriented students may be receptive to constructive feedback but discount harsh feedback. While such effects may not be unique to the feedback provided by Crowd raters, internal raters like SPs and faculty often have an opportunity to meet with students face-to-face

and manage feedback receptivity, whereas crowd ratings would be delivered with no room for the rater and ratee to come to a co-constructed understanding of the performance episode.

Finally, future studies should attempt to replicate these results using performance episodes of actual students. Although this study attempted to represent a realistic performance continuum through standardized student scripts that modeled behavior across skill levels, these stimuli are a sort of "fixed effect" and don't represent the richness and nuance of real student behavior. In particular, several Faculty raters noted that the performance of the low skill standardized students was extraordinarily poor and unrepresentative of even the lowest-performing students they encounter at RMC. Using real student data would allow us to treat student performance as a "random effect" and perhaps generalize findings to other students and performance episodes. Furthermore, the differential effects of student skill level on rating accuracy and rater perceptions of efficacy may disappear or be tempered when using a set of real videos.

Implications and Conclusion

This dissertation began by questioning the mere applicability of the gold-standard or expert model to the evaluation of medical student OSCE performance. Should rater accuracy and reliability be the primary goal of the clinical skills evaluation task? If, despite training, finely tuned instruments, and optimized evaluation conditions, we can't get raters to agree, should we interpret their disagreement as signal rather than noise? Can these goals that seem to be diametrically opposed coexist in a single measurement system, and should we value one over another? The answer, of course, is that these questions require deep reflection each time educators build, deploy, and interpret an assessment of clinical skills.

The key is to ask insightful questions and to craft an argument about the validity of the ratings, keeping in mind the *purpose and use* of the assessment. For example, does it matter if the evaluations diverge if they come from real patients evaluating future practitioner interpersonal aptitude? Probably yes, but the divergence is likely an asset, because students value these consumer perspectives as they reflect real differences in patient preference. Does it matter if the evaluations diverge with respect to whether or not the student washed her hands before performing a procedure or conducting an exam? Probably yes, but the divergence is a likely a liability since not performing this discrete behavior can lead to acute negative patient outcomes.

It is also important to consider the practical implications of this study and the viability of using a crowdsourced rating system in an operational environment. It is clear that both Crowd raters and students prefer that a crowdsourced system be used to evaluate interpersonal and communication skills over physical exam skills; Crowd raters feel less efficacious about rating PE skills and students don't view Crowd raters as a credible source for PE feedback, which means they are unlikely to engage in modifying their behavior. Students also want as much feedback as possible from as many sources as possible (but especially from Faculty) returned to them as quickly as possible. It is logistically difficult to collect ratings from SPs or Faculty at scale due to cost constraints, a limited rater pool, and the long length of time between the student's performance and the provision of feedback. Used as a complement, students would value Crowd ratings beyond internal ratings alone to improve their interpersonal and communication skills. Although such a system would add cost to an existing system, it also brings immense value by allowing students rapid insight about how a sizable cohort of real patients perceive their behavior. Perhaps then an ideal system might involve the use of crowd ratings at the faculty-student interface, where the faculty member and student review the

videotaped performance episode together in face-to-face feedback session scheduled shortly after the standardized patient encounter and use the crowd ratings and feedback to co-construct a learning plan.

References

AAMC. (2015). Physician supply and demand through 2025: Key findings. Retrieved December 6, 2015, from https://www.aamc.org/newsroom/newsreleases/426166/20150303.html

Agency for Healthcare Research and Quality. (2014). Medical Expenditure Panel Survey. Retrieved February 23, 2017, from https://meps.ahrq.gov/mepsweb/

Aghdasi, N., Bly, R., White, L. W., Hannaford, B., Moe, K., & Lendvay, T. S. (2015). Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. *Journal of Surgical Research*, *196*(2), 302–306. http://doi.org/10.1016/j.jss.2015.03.018

Alliger, G. M., Tamnenbaum, S. I., Bennett, W. J., Traver, H., & Shotland, A. (1998). A Meta-Analysis of the Relations among Training Criteria.

American Educational Research Association, American Psychological Association, National Council of Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant Reactions in Selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment*, *18*(3), 291–304. http://doi.org/10.1111/j.1468-2389.2010.00512.x

Angkaw, A. C., Tran, G. Q., & Haaga, D. A. F. (2006). Effects of training intensity on observers' ratings of anxiety, social skills, and alcohol-specific coping skills. *Behaviour Research and Therapy*, *44*(4), 533–544. http://doi.org/10.1016/j.brat.2005.04.002

Applied. (n.d.). Applied. Retrieved February 23, 2017, from https://www.beapplied.com/how-it-works

Arthur, W., Jr, & Day, E. A. (2011). Assessment centers.

Askin, E., & Moore, N. (2014). The Health Care Handbook.

Ben Christensen. (2015, November 17). Curating the crowd: When to use curated crowds vs. crowdsourcing. Retrieved February 23, 2017, from http://appen.com/7613/managing-the-crowd-when-to-use-managed-crowds-vs-crowdsourcing/

Berger, A. J., Gillespie, C. C., Tewksbury, L. R., Overstreet, I. M., Tsai, M. C., Kalet, A. L., & Ogilvie, J. B. (2012). Assessment of medical student clinical reasoning by "lay" vs physician raters: inter-rater reliability using a scoring guide in a multidisciplinary objective structured clinical examination. *Ajs*, *203*(1), 81–86. http://doi.org/10.1016/j.amjsurg.2011.08.003

Blume, B. D., Ford, J. K., & Baldwin, T. T. (2010). Transfer of training: A meta-analytic review. *Journal of …*. http://doi.org/10.1177/0149206309352880

Borsboom, B. (2012). *Guess Who?: A game to crowdsource the labeling of affective facial expressions is comparable to expert ratings.* (pp. 1–17).

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*(12), 1181–1189. http://doi.org/10.1111/j.1365-2923.2011.04075.x

Brett, J. F., & Atwater, L. E. (2001). 360 degree feedback: accuracy, reactions, and perceptions of usefulness. *The Journal of Applied Psychology*, *86*(5), 930–942.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (2016). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*.

http://doi.org/10.1177/109442819921004

Chen, C., White, L. W., Kowalewski, T. M., Aggarwal, R., Lintott, C., Comstock, B., et al. (2014). Crowd-sourced assessment of technical skills: A novel method to evaluate surgical performance. *The Journal of Surgical Research*, *187*(1), 65–71. http://doi.org/10.1016/j.jss.2013.09.024

Chenot, J.-F., Simmenroth-Nayda, A., Koch, A., Fischer, T., Scherer, M., Emmert, B., et al. (2007). Can student tutors act as examiners in an objective structured clinical examination? *Medical Education*, *41*(11), 1032–1038. http://doi.org/10.1111/j.1365-2923.2007.02895.x

Cohen, M. S., Freeman, J. T., & Thompson, B. (1998). Critical thinking skills in tactical decision making: A model and a training strategy. *Making Decisions Under Stress …*.

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *Journal of General Internal Medicine*, *24*(1), 74–79. http://doi.org/10.1007/s11606-008-0842-3

DeNisi, A. S., & Sonesh, S. (2011). The appraisal and management of performance at work.

Dickter, D. N., Stielstra, S., & Lineberry, M. (2015). Interrater reliability of standardized actors versus nonactors in a simulation based assessment of interprofessional collaboration. *Simulation in Healthcare: the Journal of the Society for Simulation in Healthcare*, *10*(4), 249–255. http://doi.org/10.1097/SIH.0000000000000094

Duffield, K. E., & Spencer, J. A. (2002). A survey of medical students' views about the purposes and fairness of assessment. *Medical Education*, *36*(9), 879–886.

Ellickson, M. C., & Logsdon, K. (2002). Determinants of Job Satisfaction of Municipal

Government Employees. *Public Personnel Management*, *31*(3), 343–358. http://doi.org/10.1177/009102600203100307

Englander, R., Cameron, T., Ballard, A. J., Dodge, J., Bull, J., & Aschenbrener, C. A. (2013). Toward a common taxonomy of competency domains for the health professions and competencies for physicians. *Academic Medicine : Journal of the Association of American Medical Colleges*, *88*(8), 1088–1094. http://doi.org/10.1097/ACM.0b013e31829a3b2b

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, *79*(10 Suppl), S70–81.

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine*, *15*(11), 988–994. http://doi.org/10.1111/j.1553-2712.2008.00227.x

Ericsson, K. A., & Smith, J. (1991). Toward a General Theory of Expertise. Cambridge University Press.

Estellés-Arolas, E. (2012). Towards an integrated crowdsourcing definition. *Journal of Information …*. http://doi.org/10.1177/016555150000000

Frank, J. R., Snell, L. S., Cate, O. T., Holmboe, E. S., Carraccio, C., Swing, S. R., et al. (2010). Competency-based medical education: theory to practice. *Medical Teacher*, *32*(8), 638–645. http://doi.org/10.3109/0142159X.2010.501190

Freelon, D. (n.d.). ReCal3. Retrieved February 23, 2017, from http://dfreelon.org/utils/recalfront/recal3/

Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature*.

Govaerts, M. J. B., Van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2002). Optimising the reproducibility of a performance-based assessment test in midwifery education. *Advances in Health Sciences Education : Theory and Practice*, *7*(2), 133–145. http://doi.org/10.1023/A:1015720302925

Harden, R. M., & Gleeson, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education*, *13*(1), 39–54. http://doi.org/10.1111/j.1365-2923.1979.tb00918.x

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant Reactions to Selection Procedures: An Updated Model and Meta-Analysis. *Personnel Psychology*, *57*(3), 639–683. http://doi.org/10.1111/j.1744-6570.2004.00003.x

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and …*. http://doi.org/10.1080/19312450709336664

Hodder, R. V., Rivington, R. N., Calcutt, L. E., & Hart, I. R. (1989). The effectiveness of immediate feedback during the Objective Structured Clinical Examination. *Medical Education*, *23*(2), 184–188. http://doi.org/10.1111/j.1365-2923.1989.tb00884.x

Holmboe, E. S. (2015). Realizing the promise of competency-based medical education. *Academic Medicine*, *90*(4), 411–413. http://doi.org/10.1097/ACM.0000000000000515

Holmboe, E. S., Huot, S., Chung, J., Norcini, J., & Hawkins, R. E. (2003). Construct validity of the miniclinical evaluation exercise (miniCEX). *Academic Medicine*, *78*(8), 826–830.

Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., Frank, J. R., & Collaborators, F. T. I. C. (2010). The role of assessment in competency-based medical education. *Medical Teacher*, *32*(8), 676–682. http://doi.org/10.3109/0142159X.2010.500704

Holst, D., Kowalewski, T. M., White, L. W., Brand, T. C., Harper, J. D., Sorensen, M. D., Kirsch, S., et al. (2015a). Crowd-sourced assessment of technical skills: an adjunct to urology resident surgical simulation training. *Journal of Endourology*, *29*(5), 604–609. http://doi.org/10.1089/end.2014.0616

Holst, D., Kowalewski, T. M., White, L. W., Brand, T. C., Harper, J. D., Sorensen, M. D., Truong, M., et al. (2015b). Crowd-sourced assessment of technical skills: Differentiating animate surgical skill through the wisdom of crowds. *Journal of Endourology*, *29*(10), 1183–1188. http://doi.org/10.1089/end.2015.0104

Horvath, M., & Andrews, S. B. (2007). The role of fairness perceptions and accountability attributions in predicting reactions to organizational events. *The Journal of Psychology*, *141*(2), 203–222. http://doi.org/10.3200/JRLP.141.2.203-223

Ilgen, J. S., Ma, I. W. Y., Hatala, R., & Cook, D. A. (2015). A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Medical Education*, *49*(2), 161–173. http://doi.org/10.1111/medu.12621

Kalet, D. A., Earp, J. A., & Kowlowitz, V. (1992). How well do faculty evaluate the interviewing skills of medical students? *Journal of General Internal Medicine*, *7*(5), 499–505. http://doi.org/10.1007/BF02599452

Kane, M. T. (1992). The Assessment of Professional Competence. *Evaluation & the Health Professions*, *15*(2), 163–182. http://doi.org/10.1177/016327879201500203

Khatib, F., DiMaio, F., Foldit Contenders Group, Foldit Void Crushers Group, Cooper, S., Kazmierczyk, M., et al. (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural & Molecular Biology*, *18*(10), 1175–1177. http://doi.org/10.1038/nsmb.2119

Ladyshewsky, R., Baker, R., Jones, M., & Nelson, L. (2000). Reliability and validity of an extended simulated patient case: A tool for evaluation and research in physiotherapy. *Physiotherapy Theory and Practice*, *16*(1), 15–25. http://doi.org/10.1080/095939800307575

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72–107. http://doi.org/10.1037/0033-2909.87.1.72

Lang, F., McCord, R., Harvill, L., & Anderson, D. S. (2004). Communication assessment using the common ground instrument: psychometric properties. *Family Medicine*, *36*(3), 189–198.

Legree, P. J., Psotka, J., Tremble, T. R., Jr, & Bourne, D. (2005). Applying consensus based measurement to the assessment of emerging domains. *Leadership & Organization Development Journal* (Vol. 36, pp. 120–136). http://doi.org/10.1108/LODJ-01-2013-0008

Mazor, K. M., Ockene, J. K., Rogers, H. J., Carlin, M. M., & Quirk, M. E. (2005). The relationship between checklist scores on a communication OSCE and analogue patients' perceptions of communication. *Advances in Health Sciences Education : Theory and Practice*, *10*(1), 37–51. http://doi.org/10.1007/s10459-004-1790-2

Mazor, K. M., Zanetti, M. L., Alper, E. J., Hatem, D., Barrett, S. V., Meterko, V., et al. (2007). Assessing professionalism in the context of an objective structured clinical examination: an in-depth study of the rating process. *Medical Education*, *41*(4), 331–340. http://doi.org/10.1111/j.1365-2929.2006.02692.x

Michel, A. (2016). Harnessing the wisdom of crowds to improve hiring. *APS Observer*, *30*(1).

Moineau, G., Power, B., Pion, A.-M. J., Wood, T. J., & Humphrey-Murto, S. (2010).

Comparison of student examiner to faculty examiner scoring and feedback in an OSCE. *Medical Education*, *45*(2), 183–191. http://doi.org/10.1111/j.1365-2923.2010.03800.x

Müller, M. J., & Dragicevic, A. (2003). Standardized rater training for the Hamilton Depression Rating Scale (HAMD-17) in psychiatric novices. *Journal of Affective Disorders*, *77*(1), 65–69. http://doi.org/10.1016/S0165-0327(02)00097-6

Müller, M. J., Rossbach, W., Dannigkeit, P., Müller-Siecheneder, F., Szegedi, A., & Wetzel, H. (1998). Evaluation of standardized rater training for the Positive and Negative Syndrome Scale (PANSS). *Schizophrenia Research*, *32*(3), 151–160. http://doi.org/10.1016/S0920-9964(98)00051-6

Newble, D. I., Hoare, J., & Sheldrake, P. F. (1980). The selection and training of examiners for clinical examinations. *Medical Education*, *14*(5), 345–349. http://doi.org/10.1111/j.1365-2923.1980.tb02379.x

Newlin-Canzone, E. T., Scerbo, M. W., Gliva-McConvey, G., & Wallace, A. M. (2013). The cognitive demands of standardized patients. *Simulation in Healthcare: the Journal of the Society for Simulation in Healthcare*, *8*(4), 207–214. http://doi.org/10.1097/SIH.0b013e31828b419e

Noel, G. L., Herbers, J. E., Caplow, M. P., Cooper, G. S., Pangaro, L. N., & Harvey, J. (1992). How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine*, *117*(9), 757–765.

Norman, G. (2005). Editorial--checklists vs. ratings, the illusion of objectivity, the demise of skills and the debasement of evidence. *Advances in Health Sciences Education : Theory and Practice*, *10*(1), 1–3. http://doi.org/10.1007/s10459-005-4723-9

Noronha, J., Hysen, E., Zhang, H., & Gajos, K. Z. (2011). Platemate: Crowdsourcing nutritional

analysis from food photographs. *Proceedings of the 24th …* (pp. 1–12). ACM. http://doi.org/10.1145/2047196.2047198

Pangaro, L. N., & McGaghie, W. C. (2015). Handbook on Medical Student Evaluation and Assessment.

Parsons, D. (2011, December). US government turns to crowdsourcing for intelligence. Retrieved February 23, 2017, from http://www.nationaldefensemagazine.org/archive/2011/december/pages/usgovernmentturnstocrowdsourcingforintelligence.aspx

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

Sackett, P. R., & Wilson, M. A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. *Journal of Applied Psychology*, *67*(1), 10–17. http://doi.org/10.1037/0021-9010.67.1.10

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. (1990). A cognitive perspective on medical expertise: theory and implication [published erratum appears in Acad Med 1992 Apr;67(4):287]. *Academic Medicine*, *65*(10), 611–11. http://doi.org/10.1097/ACM.0b013e3181e9148e

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory. Newbury Park, CA: Sage Publications.

Sheehan, K. B., & Pittman, M. (2016). Amazon's Mechanical Turk for academics. Irvine, CA: Melvin & Leigh.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and

training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27–33. http://doi.org/10.1111/j.1540-4781.1992.tb02574.x

Shulman, L. S. (1970). Cognitive learning and the educational process. *Academic Medicine*, *45*(11), 90.

Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *The Journal of Applied Psychology*, *93*(2), 280–295. http://doi.org/10.1037/0021-9010.93.2.280

Stilson, F. (2009). Psychometrics of OSCE standardized patient measurements.

Sturpe, D. A., Huynh, D., & Haines, S. T. (2010). Scoring objective structured clinical examinations using video monitors or video recordings. *American Journal of …*.

Sun, L. (n.d.). Demand is high for pretend patients.

Surowiecki, J. (2005). The Wisdom of Crowds. New York, NY: Anchor.

Turner-McGrievy, G., Helander, E. E., Kaipainen, K., Perez-Macias, J. M., & Korhonen, I. (2014). The use of crowdsourcing for dietary self-monitoring: Crowdsourced ratings of food pictures are comparable to trained observers. *Journal of the American Medical Informatics Association*, *22*(e1), 1–6. http://doi.org/10.1136/amiajnl-2014-002636

US Bureau of Labor Statistics. (2016, December 8). Employer costs for employee compensation. Retrieved February 23, 2017, from https://www.bls.gov/news.release/pdf/ecec.pdf

US Bureau of Labor Statistics. (n.d.). Occupational employment statistics. Retrieved February 23, 2017, from https://www.bls.gov/OES/

US Census Bureau, US Bureau of Labor Statistics. (n.d.). Current Population Survey (2015). Retrieved February 23, 2017, from http://www.census.gov/programs-surveys/cps.html

USMLE content outline. (n.d.). USMLE content outline. Retrieved February 23, 2017, from

http://www.usmle.org/pdfs/usmlecontentoutline.pdf

Waples, C. J. (2015). Receptivity to feedback: an investigation of the influence of feedback sign, feedback specificity, and goal orientation.

Weekley, J. A., & Gier, J. A. (1989). Ceilings in the Reliability and Validity of Performance Ratings: The Case of Expert Raters. *Academy of Management Journal*, *32*(1), 213–222. http://doi.org/10.2307/256428

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology*.

Wenrich, M. D., Carline, J. D., Giles, L. M., & Ramsey, P. G. (1993). Ratings of the performances of practicing internists by hospital-based registered nurses. *Academic Medicine*, *68*(9), 680–687.

White, L. W., Kowalewski, T. M., Dockter, R. L., Comstock, B., Hannaford, B., & Lendvay, T. S. (2015). Crowd-sourced assessment of technical skill: A valid method for discriminating basic robotic surgery skills. *Journal of Endourology*, *29*(11), 1295–1301. http://doi.org/10.1089/end.2015.0191

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine*, *15*(4), 270–292. http://doi.org/10.1207/S15328015TLM1504_11

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189–205. http://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Zanetti, M. L., Keller, L., Mazor, K., Carlin, M. M., Alper, E. J., Hatem, D., et al. (2010). Using standardized patients to assess professionalism: A generalizability study. *Teaching and Learning in Medicine*, *22*(4), 274–279. http://doi.org/10.1080/10401334.2010.512542

APPENDICES

Appendix 1: IRB Approval Letters

USF
UNIVERSITY OF
SOUTH FLORIDA

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX(813)974-7091

November 2, 2016

Mark Grichanik
Psychology
Tampa, FL 33612

RE: **Exempt Certification**
IRB#: Pro00028287
Title: Many hands make light work: Crowdsourced ratings of medical student OSCE performance

Dear Mr. Grichanik:

On 11/2/2016, the Institutional Review Board (IRB) determined that your research meets criteria for exemption from the federal regulations as outlined by 45CFR46.101(b):

(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless:
(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

As the principal investigator for this study, it is your responsibility to ensure that this research is conducted as outlined in your application and consistent with the ethical principles outlined in the Belmont Report and with USF HRPP policies and procedures.

Please note, as per USF HRPP Policy, once the Exempt determination is made, the application is closed in ARC. Any proposed or anticipated changes to the study design that was previously declared exempt from IRB review must be submitted to the IRB as a new study prior to initiation of the change. However, administrative changes, including changes in research personnel, do not warrant an amendment or new application.

Given the determination of exemption, this application is being closed in ARC. This does not limit your ability to conduct your research project.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have

any questions regarding this matter, please call 813-974-5638.

Sincerely,

John A Schinka, Ph.D.

John Schinka, Ph.D., Chairperson
USF Institutional Review Board

USF
UNIVERSITY OF
SOUTH FLORIDA

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX(813)974-7091

December 6, 2016

Mark Grichanik
Psychology
Tampa, FL 33612

RE: **Exempt Certification**
IRB#: Pro00028648
Title: Student Reactions to Crowdsourced OSCE Ratings

Dear Mr. Grichanik:

On 12/5/2016, the Institutional Review Board (IRB) determined that your research meets criteria for exemption from the federal regulations as outlined by 45CFR46.101(b):

(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless:
(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

As the principal investigator for this study, it is your responsibility to ensure that this research is conducted as outlined in your application and consistent with the ethical principles outlined in the Belmont Report and with USF HRPP policies and procedures.

Please note, as per USF HRPP Policy, once the Exempt determination is made, the application is closed in ARC. Any proposed or anticipated changes to the study design that was previously declared exempt from IRB review must be submitted to the IRB as a new study prior to initiation of the change. However, administrative changes, including changes in research personnel, do not warrant an amendment or new application.

Given the determination of exemption, this application is being closed in ARC. This does not limit your ability to conduct your research project.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

John A Schinka, Ph.D.

John Schinka, Ph.D., Chairperson
USF Institutional Review Board

RUSH UNIVERSITY MEDICAL CENTER
1653 WEST CONGRESS PARKWAY, CHICAGO, ILLINOIS, 60612-3833
RUSH UNIVERSITY

RUSH UNIVERSITY MEDICAL CENTER

Full Accreditation

OFFICE OF RESEARCH AFFAIRS
312.942.5498
312.942.2874 (FAX)

Institutional Review Board #1
FWA #: 00000482

Notification of Exemption from IRB Review

To: Mark Grichanik
ORA #: 16100401-IRB01
Project Title: Crowdsourced OSCE Ratings

Date Exemption Granted: 11/4/2016

Dear Mark Grichanik ,

This exemption was granted for the following reasons:

Question Text	Part Number
This research will be conducted in established or commonly accepted educational settings, involving normal educational practices, such as research on regular and special education instructional strategies, or research on the effectiveness of instructional techniques, curricula, or classroom management methods.	45CFR46.101(b)(1)
This research involves the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior. Information obtained is recorded in such a manner that human subjects cannot be identified, directly or through identifiers linked to the subjects; and any disclosure of the human subjects' responses outside the research does not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.	45CFR46.101(b)(2)

If you change your protocol in any way, these issues must be re-reviewed.

Please note that if your study is retrospective that only data collected before 11/4/2016 may be analyzed and not after that date.

Our Institutional Federwide Assurance Number is FWA00000482.

Thank you for your submission. Good luck with your project.

{The below is a representation of an electronic record that was signed electronically and is the manifestation of the electronic signature.}

John Cobb
11/4/2016 1:44 PM
Signing for Mary Jane Welch

Mary Jane Welch, DNP, APRN, BC, CIP
Rush University Medical Center
Assoc. VP, Research Regulatory Operations
Associate Professor, College of Nursing

RUSH UNIVERSITY MEDICAL CENTER
1653 WEST CONGRESS PARKWAY, CHICAGO, ILLINOIS, 60612-3833
RUSH UNIVERSITY

RUSH UNIVERSITY MEDICAL CENTER

Full Accreditation

OFFICE OF RESEARCH AFFAIRS
312.942.5498
312.942.2874 (FAX)

Institutional Review Board #2
FWA #: 00000482

Notification of Exemption from IRB Review

To: Mark Grichanik
ORA #: 16111503-IRB02
Project Title: Student Reactions to Crowdsourced OSCE Ratings

Date Exemption Granted: 12/12/2016

Dear Mark Grichanik ,

This exemption was granted for the following reasons:

Question Text	Part Number
This research involves the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior. Information obtained is recorded in such a manner that human subjects cannot be identified, directly or through identifiers linked to the subjects; and any disclosure of the human subjects' responses outside the research does not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.	45CFR46.101(b)(2)

If you change your protocol in any way, these issues must be re-reviewed.

Please note that if your study is retrospective that only data collected before may be analyzed and not after that date.

Our Institutional Federalwide Assurance Number is FWA00000482.

Thank you for your submission. Good luck with your project.

{The below is a representation of an electronic record that was signed electronically and is the manifestation of the electronic signature.}

John Cobb
12/12/2016 9:31 AM
Signing for Mary Jane Welch

Mary Jane Welch, DNP, APRN, BC, CIP
Rush University Medical Center
Assoc. VP, Research Regulatory Operations
Associate Professor, College of Nursing

Appendix 2: Interpersonal and Communication Skills Questionnaire (ICSF)

1) Beginning the Encounter: Student ability to make introductions and initiate discussion with clarity, ease, and confidence.

	Yes	No
The student addressed the patient respectfully using the patient's last name (ICS_1_Chx)	( )	( )
The student introduced him/herself by name (ICS_2_Chx)	( )	( )
The student described his/her role on the healthcare team (ICS_3_Chx)	( )	( )

2) Gathering Information About Patient's Concerns: Student ability to collect information about the patient's presenting concern, relevant aspects of the patient's personal and medical history, and the patient's perspective on his/her concern. (ICS_4)

1. ( ) The student did not ask about or spent little time on the patient's concerns, missed significant details about the patient's concerns, **OR** framed questions in a way that did not allow the patient to tell his/her story.
2. ( )
3. ( ) The student asked about the patient's concern and allowed the patient adequate time to share his/her story. The student got the essential information. There may have been some missed opportunities to gain a deeper understanding of the patient's concern.
4. ( )
5. ( ) The student allowed the patient to fully share his/her story in a conversational and comfortable way. The student got all of the essential information and all or almost all of the details surrounding the patient's concerns.

3) Empathy for the Patient's Distress: Student ability to 1) recognize the patient's concern and emotion, and 2) respond to the patient's emotion in an accurate, compassionate, and genuine manner. (ICS_5)

1. ( ) The student never asked about the patient's perspective of his/her concerns or illness, never asked about the patient's expectations for treatment, **OR** did not understand how these concerns impact the patient emotionally or functionally.
2. ( )
3. ( ) The student made an effort to understand the patient's perspective of his/her concerns, his/her expectations for treatment, and how these concerns impact the patient emotionally or functionally.
4. ( )
5. ( ) The student thoroughly explored and demonstrated understanding of the patient's perspective on his/her concerns including his/her expectations for treatment and how these concerns impact the patient emotionally and functionally.

4) Providing Patient with Information: Student ability to 1) share information with the patient, and 2) respond to the patient's questions regarding his/her concern and possible diagnosis. (ICS_6)

1. ( ) The student did not encourage the patient to ask questions. The student rarely provided the patient with information about his/her concern. The student did not help the patient understand his/her symptoms or possible diagnosis.
2. ( )
3. ( ) The student answered questions but **DID NOT** check for the patient's understanding. The student provided the patient with an appropriate amount and complexity of information.
4. ( )
5. ( ) The student clearly answered the patient's questions. The student provided the patient with clear information about his/her concern at a level that was easy for the patient to understand. The student checked to be sure the patient understood.

5) Using Appropriate & Sensitive Language: Student ability to 1) use medical language in a way that is neither too technical nor too simplistic, 2) avoid presenting information in an unnecessarily frightening manner, and 3) appropriately defer questions to a supervising physician when he/she cannot answer the patient's questions. (ICS_7)

1. ( ) The student frequently used medical jargon or frequently made frightening or alarming comments. The student addressed sensitive topics in an abrupt manner, or in a manner which expressed judgment.
2. ( )
3. ( ) The student occasionally used medical jargon without explaining it in layman's terms. The student rarely made unduly frightening comments. The student addressed sensitive topics professionally but in a detached manner.
4. ( )
5. ( ) The student used medical terms as needed, and explained them in a way that the patient could understand without the patient asking for clarification. The student explained serious concerns and findings in a clear but compassionate way. The student addressed sensitive topics in a way that made the patient feel at ease.

6) Making Decisions: Student ability to 1) provide education about treatment options, 2) solicit the patient's opinion on treatment options, and 3) develop a treatment plan in a collaborative manner. (ICS_8)

1. ( ) The student did not share information on different treatment options. The student did not include the patient in decision-making. The student did not assess the patient's willingness to execute the plan.
2. ( )
3. The student presented the patient with information about different treatment options, but did not assess the patient's preference for treatment planning.
4. ( )
5. The student presented the patient with different treatment options, provided rationale for each, and assessed the patient's treatment preference. The student helped the patient choose a treatment option in a collaborative way.

7) Supporting Emotions & Fostering Relationship: Learner ability to 1) connect with the patient as a person through eye contact and body language, and 2) demonstrate an interest in the patient beyond the "facts" of the patient's concern. (ICS_9)

1. ( ) The student did not maintain eye contact or the student's eye contact was uncomfortable. The student's body language was closed or off-putting. The student rarely reflected the patient's emotions. The student did not demonstrate genuine interest in the patient as a person.
2. ( )
3. The student maintained eye contact. The student's body language was open. The student reflected the patient's emotions a few times in the encounter. The student was pleasant but distant. The student did not demonstrate genuine interest in the patient as a person.
4. ( )
5. The student's body language and demeanor was warm and inviting. The student frequently reflected the patient's emotions. The student and patient connected well, and the student demonstrated genuine interest in the patient as a person.

8) Ending the Encounter: Student ability to 1) confidently and cohesively summarize the encounter, 2) provide an opportunity for me to ask remaining questions, and 3) provide instructions for next steps.

	Yes	**No**
The student summarized the patient's reason for visit. (ICS_10_Chx)	( )	( )
The student checked for accuracy of their summary. (ICS_11_Chx)	( )	( )
The student provided the patient with instructions on what to do after he/she left the room. (ICS_12_Chx)	( )	( )

9) Overall Interpersonal and Communication Skills: **Overall assessment of student's interpersonal and communication skill.** (ICS_13)

1. ( ) The student was difficult to talk with, the patient felt uncomfortable, **OR** the encounter was disorganized or confusing.
2. ( )
3. ( ) The student was pleasant to talk with, and the patient rarely felt uncomfortable. The student's organization was variable, but the patient generally understood the student's line of questioning.
4. ( )
5. ( ) The student was easy to talk with. The patient felt comfortable and respected throughout the encounter. The student's organization was excellent such that the student and patient were understanding one another and collaborating throughout the encounter.

10) Probability of Return: **Based on observing this interaction, how likely would you be to seek this provider's care as a patient or recommend this provider to family and friends?** (ICS_14)

1. ( ) I would not see this student as a patient
2. ( )
3. ( ) I would see this student as a patient
4. ( )
5. ( ) I would see this student as a patient and recommend him/her as a physician to family or friends

11) With respect to interpersonal and communication skills, what did this student DO WELL **in this patient encounter?**

_____
_____
_____
_____

12) With respect to interpersonal and communication skills, what could this student DO BETTER **in his/her next patient encounter?**

_____
_____
_____
_____

Appendix 3: Physical Exam Skills Checklist (PESC)

Visible Done Correctly = You saw clearly that the student performed this maneuver correctly
Visible Done Incorrectly = You saw clearly that the student performed this maneuver incorrectly
Visible Not Done = You saw clearly that the student did not perform this maneuver
Observation Obscured = The video production obscures a critical portion of this maneuver such that it is difficult for you to be sure whether the student performed the maneuver correctly or not

	Visible **Done Correctly**	Visible **Done Incorrectly**	Visible **Not Done**	**Observation Obscured**
1. Clean hands before touching the patient (using hand sanitizer or soap and water) PE_1	( )	( )	( )	( )
2. Visually inspect skin on chest (from collarbone to bra line). Inform patient of visual inspection. PE_2	( )	( )	( )	( )
3. Place finger pads over aortic area to palpate for any thrills or pulsations PE_3	( )	( )	( )	( )
4. Place finger pads over pulmonic area to palpate for any thrills or pulsations PE_4	( )	( )	( )	( )
5. Place ulnar surface of the palm over the tricuspid area to palpate for any thrills or pulsations PE_5	( )	( )	( )	( )
6. Place finger pads over the mitral area of the heart to feel the point of maximum impulse PE_6	( )	( )	( )	( )
7. Listen with diaphragm of stethoscope for heart sounds	( )	( )	( )	( )

in the aortic area PE_7				
8. Listen with diaphragm of stethoscope for heart sounds in the pulmonic area PE_8	( )	( )	( )	( )
9. Listen with diaphragm of stethoscope for heart sounds in the tricuspid area PE_9	( )	( )	( )	( )
10. Listen with diaphragm of stethoscope for heart sounds in the mitral area PE_10	( )	( )	( )	( )
11. Listen with bell of stethoscope for heart sounds in the aortic area PE_11	( )	( )	( )	( )
12. Listen with bell of stethoscope for heart sounds in the pulmonic area PE_12	( )	( )	( )	( )
13. Listen with bell of stethoscope for heart sounds in the tricuspid area PE_13	( )	( )	( )	( )
14. Listen with bell of stethoscope for heart sounds in the mitral area PE_14	( )	( )	( )	( )
15. With the patient lying on his/her left side, listen with bell of stethoscope for heart sounds in the mitral area PE_15	( )	( )	( )	( )
16. Stands to the patient's right during the entire exam PE_16	( )	( )	( )	( )

Please justify all maneuvers marked as "Done Incorrectly". List each maneuver by number and explain what was done incorrectly.

For example: "#15 - Patient was lying on right side instead of left side"

_____

_____

Please justify all maneuvers marked as "Observation Obscured". List each maneuver by number and explain the obstruction/source of ambiguity.

For example: "#1 - Saw student rubbing hands together at the beginning of the video, but did not directly see student applying hand soap or sanitizer"

_____

_____

Appendix 4: Interpersonal and Communication Skills Task Crowd Rater Reactions Questionnaire (RRX-ICSF)

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
I felt comfortable evaluating this student's social and communication skills (RRX_1)	( )	( )	( )	( )	( )	( )
I felt comfortable providing positive feedback about the student's social and communication skills (RRX_2)	( )	( )	( )	( )	( )	( )
I felt comfortable providing negative feedback about the student's social and communication skills (RRX_3)	( )	( )	( )	( )	( )	( )
I understood the elements of the questionnaire well (RRX_4)	( )	( )	( )	( )	( )	( )
The questionnaire gave me enough information to make decisions about the student's social and communication skills (RRX_5)	( )	( )	( )	( )	( )	( )
The instructions for completing the questionnaire were clear (RRX_6)	( )	( )	( )	( )	( )	( )
I was motivated to make accurate ratings of this student's social and communication skills (RRX_7)	( )	( )	( )	( )	( )	( )
I was motivated to be fair in evaluating the student's social and communication skills (RRX_8)	( )	( )	( )	( )	( )	( )
Patients, like myself, should be able to provide feedback about the social and communication skills of future physicians (RRX_9)	( )	( )	( )	( )	( )	( )

Patients, like myself, are capable of providing feedback about the social and communication skills of future physicians (RRX_10)	( )	( )	( )	( )	( )	( )
I feel like I need additional training to make such evaluations in the future (RRX_11)	( )	( )	( )	( )	( )	( )

Please provide comments here regarding <u>your experience</u> evaluating this encounter. For example, you may comment on the instructions you received, the evaluation tool itself, your feelings about evaluating medical student performance, etc.

_____
_____
_____
_____

Appendix 5: Physical Exam Task Crowd Rater Reactions Questionnaire (RRX-PESC)

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
I felt comfortable evaluating this student's physical exam skills (RRX_1)	( )	( )	( )	( )	( )	( )
I understood the elements of the questionnaire well (RRX_4)	( )	( )	( )	( )	( )	( )
The questionnaire gave me enough information to make decisions about the student's physical exam skills (RRX_5)	( )	( )	( )	( )	( )	( )
The instructions for completing the questionnaire were clear (RRX_6)	( )	( )	( )	( )	( )	( )
I was motivated to make accurate ratings of this student's physical exam skills (RRX_7)	( )	( )	( )	( )	( )	( )
I was motivated to be fair in evaluating the student's physical exam skills (RRX_8)	( )	( )	( )	( )	( )	( )
Patients, like myself, should be able to provide feedback about the physical exam skills of future physicians (RRX_9)	( )	( )	( )	( )	( )	( )
Patients, like myself, are capable of providing feedback about the	( )	( )	( )	( )	( )	( )

physical exam skills of future physicians (RRX_10)						
I feel like I need additional training to make such evaluations in the future (RRX_11)	( )	( )	( )	( )	( )	( )

Please provide comments here regarding <u>your experience</u> evaluating this encounter. For example, you may comment on the instructions you received, the evaluation tool itself, your feelings about evaluating medical student performance, etc.

_____
_____
_____
_____

Appendix 6: Demographics Questions by Participant Type

BhvFac:

1. Approximately how many years have you been teaching in the Rush Medical College Patient Interviewing Course (formerly known as the Interviewing and Communication Course)?

2. Approximately how many student videos have you evaluated for interpersonal/social/communication skills as part of your teaching in the Rush Medical College Patient Interviewing Course (formerly known as the Interviewing and Communication Course)?

ClinFac:

1. Approximately how many years have you practiced medicine at any institution?

2. Do you teach Physical Diagnosis at Rush Medical College?
   ( ) Yes
   ( ) No

3. Approximately how many years have you taught Physical Diagnosis at Rush Medical College?

4. In your lifetime, approximately how many students, residents, or other healthcare professionals have you **taught** to perform a cardiovascular exam?

5. In your lifetime, approximately how many students, residents, or other healthcare professionals have you **evaluated** performing a cardiovascular exam?

SP:

SP EXPERIENCE AT RUSH MEDICAL COLLEGE

1. Approximately how many years have you worked as an SP at Rush Medical College?

2. Approximately how many one-on-one student encounters have you participated in as a standardized patient Rush Medical College?

3. Approximately what percentage of all encounters at Rush Medical College involved evaluating interpersonal/social/communication skills?

4. Approximately what percentage of all encounters at Rush Medical College involved evaluating physical exam skills?

5. Approximately what percentage of all encounters at Rush Medical College involved evaluating cardiovascular exam skills?

SP EXPERIENCE AT ANY INSTITUTION

6. Approximately how many years have you worked as an SP at any institution?

7. Approximately how many one-on-one student encounters have you participated in as a standardized patient at any institution?

8. Approximately what percentage of all encounters at any institution involved evaluating interpersonal/social/communication skills?

9. Approximately what percentage of all encounters at any institution involved evaluating physical exam skills?

10. Approximately what percentage of all encounters at any institution involved evaluating cardiovascular exam skills?

11. Approximately what percentage of all encounters at any institution involved evaluating the encounter using a video review?

12. Have you ever formally served as an SP trainer (i.e., trained other SPs) in evaluating the following skills?

	Yes	**No**
Interpersonal/social/communication skills	( )	( )
Physical exam skills	( )	( )
Cardiovascular exam skills	( )	( )

Crowd:

1. Age _____

2. Race
   ( ) Asian
   ( ) Native Hawaiian or Other Pacific Islander
   ( ) Black/African-American
   ( ) White
   ( ) Hispanic/Latino

( ) American Indian/Alaska Native
( ) Other - Write In: _____

3. Gender
   ( ) Male
   ( ) Female
   ( ) Transgender Male
   ( ) Transgender Female
   ( ) Gender Variant / Non-conforming
   ( ) Other - Write In: _____

4. Highest level of education
   ( ) Less than high school
   ( ) Graduated high school
   ( ) Trade/technical school
   ( ) Some college, no degree
   ( ) Associate degree
   ( ) Bachelor's degree
   ( ) Advanced degree (Master's, Ph.D., M.D.)

5. Total household income
   ( ) Less than $25,000
   ( ) $25,000 to $34,999
   ( ) $35,000 to $49,999
   ( ) $50,000 to $74,999
   ( ) $75,000 to $99,999
   ( ) $100,000 to $124,999
   ( ) $125,000 to $149,999
   ( ) $150,000 or more

6. Marital status
   ( ) Married / Domestic Partner
   ( ) Widowed
   ( ) Divorced
   ( ) Separated
   ( ) Single / Never Married

7. Is your occupation related to healthcare?
   ( ) Yes
   ( ) No

8. **In what capacity is your occupation related to healthcare?**
   _____
   _____

Students

1. Age _____

2. Race
   ( ) Asian
   ( ) Native Hawaiian or Other Pacific Islander
   ( ) Black/African-American
   ( ) White
   ( ) Hispanic/Latino
   ( ) American Indian/Alaska Native
   ( ) Other - Write In: _____

3. Gender
   ( ) Male
   ( ) Female
   ( ) Transgender Male
   ( ) Transgender Female
   ( ) Gender Variant / Non-conforming
   ( ) Other - Write In: _____

Appendix 7: Sample Video Frames for ICS and PE Videos

ICS

PE

Appendix 8: Sample SP Case Door Chart

DEMOGRAPHICS	CARE TEAM AND COMMUNICATIONS
Name: J. James Age: 42 years old Vitals BP 120/80 HR 88 RR 10 Temp 97.2	Primary Care Physician: Dr. Robert Harris

PRESENTING SITUATION:
You are a 2nd year medical student working in the office of primary care physician, Dr. Robert Harris. Patient has been "roomed" by the medical assistant, Ms. Anna Jones. She collected the following vitals:

BP: 120/80
HR: 88
RR: 10
Temp: 97.2

You have been asked by the primary care physician to meet with patient. The patient is here with a complaint of chest pain.

YOUR TASKS FOR TODAY:

1. Take a history of the patient's chief complaint of chest pain.
2. Take a social history of elements related to chief complaint.
3. Perform a cardiovascular examination.
4. Provide patient with preliminary assessment and plan.

TIME LIMIT:
15 minutes (the proctor will knock on the door at 10 minutes and again at 2 minutes)

POST-ENCOUNTER ACTIVITY:
Once you finish your encounter, you are finished for today. You do NOT need to document this encounter.

Appendix 9: Survey Gizmo Layout for the Physical Exam Skills Task for Physician Faculty Raters

Page 1: Welcome + Screening

I am a clinical faculty member at Rush Medical College
( ) Yes
( ) No

Page 2: Consent

[CONSENT WAS HERE]

Page 3: Instructions

Instructions

You are about to watch a series of six videos in which medical students participate in standardized patient encounters.

1. Before watching the first video, scroll down and familiarize yourself with the questionnaire you will use to provide feedback about the student's cardiovascular exam skills.

2. Watch each video in its entirety
   - *We strongly recommend watching the video in full screen in a new browser tab*: Click the "YouTube" icon in the bottom right corner. After the video begins to play, click the frame icon in the bottom right corner to make it full screen.
   - The student's face is blurred in each video. We understand some information about the student's behavior is lost this way, but please do your best with the behavior you *are* able to observe.
   - You will see 2-3 camera angles of the same encounter presented simultaneously
   - Make sure to turn up the sound so you can hear what the student and standardized patient are saying
   - You are not allowed to copy, retain, or distribute this video or video URL
3. After watching each the video, fill out the corresponding questionnaire
   - You can re-watch any part of the video you want while filling out each questionnaire
   - Please finalize the evaluation for each student before moving on to the next student. You will not be able to navigate back to previous evaluations.
4. After filling out evaluations for all six videos, navigate to the final page to answer a few short questions about your experience as a faculty member

If you cannot finish reviewing all of the videos in a single sitting and need to come back to finish at a later time, you can press the "Save and continue later" button found at the top of the page.

Pages 4-9: Physical Exam Skills Evaluations

[PHYSICAL EXAM VIDEO WITH PESC]

Page 10: Demographics

[PHYSICIAN FACULTY RATER DEMOGRAPHICS WAS HERE]

Page 11: Thank You

Appendix 10: Survey Gizmo Layout for the Interpersonal and Communication Skills Task for Behavioral Science Faculty Raters

Page 1: Welcome + Screening

I teach Patient Interviewing at Rush Medical College
( ) Yes
( ) No

I have received training on the Rush Medical College Interpersonal and Communications Skills (ICS) Questionnaire
( ) Yes
( ) No

Page 2: Consent

[CONSENT WAS HERE]

Page 3: Instructions

Instructions

You are about to watch a series of videos in which medical students participate in standardized patient encounters.

Please rely on the training you received at Rush Medical College in using the Interpersonal and Communication Skills (ICS) Questionnaire to characterize the performance level of each of the six students.

1. Before watching the first video, scroll down and re-familiarize yourself with the questionnaires you will use to provide feedback about the student's interpersonal and communication skills.

2. Watch each video in its entirety

- *We strongly recommend watching the video in full screen in a new browser tab*: Click the "YouTube" icon in the bottom right corner. After the video begins to play, click the frame icon in the bottom right corner to make it full screen.
- The student's face is blurred in each video. We understand some information about the student's behavior is lost this way, but please do your best with the behavior you *are* able to observe.
- You will see 2-3 camera angles of the same encounter presented simultaneously
- The physical exam segment has been removed from this video, so you will see a brief cutaway/transition.
- Make sure to turn up the sound so you can hear what the student and standardized patient are saying
- You are not allowed to copy, retain, or distribute this video or video URL

3. After watching each the video, fill out the corresponding questionnaire
    - You can re-watch any part of the video you want while filling out each questionnaire
    - Please finalize the evaluation for each student before moving on to the next student. You will not be able to navigate back to previous evaluations.
4. After filling out evaluations for all six videos, navigate to the final page to answer a few short questions about your experience as a faculty member

If you cannot finish reviewing all of the videos in a single sitting and need to come back to finish at a later time, you can press the "Save and continue later" button found at the top of the page.

Pages 4-9: Interpersonal and Communications Skills Evaluations

[INTERVIEW AND PLAN VIDEO WITH ICSF]

Page 10: Demographics

[BEHAVIORAL SCIENCE FACULTY RATER DEMOGRAPHICS WAS HERE]

Page 11: Thank You

Appendix 11: Survey Gizmo Layout for the Interpersonal and Communication Skills and Physical Exam Skills Tasks for Standardized Patient Raters

Page 1: Welcome + Screening

I am a standardized patient at Rush Medical College
( ) Yes
( ) No

I have received training on the Rush Medical College Interpersonal and Communications Skills (ICS) Questionnaire
( ) Yes
( ) No

I have received training on the Rush Medical College Cardiovascular Exam Checklist
( ) Yes
( ) No

Page 2: Consent

[CONSENT WAS HERE]

Page 3: Instructions

Instructions

You are about to watch a series of videos in which actors play the role of medical students ("standardized student") participating in standardized patient encounters.

Please rely on the training you received as a standardized patient at Rush Medical College in using the Interpersonal and Communication Skills (ICS) Questionnaire and the Cardiovascular Exam Checklist to characterize the performance level of the student in each video. In total, you will evaluate 6 interview/plan video segments and 6 physical exam video segments.

1. Before watching the first video, scroll down and re-familiarize yourself with the questionnaires you will use to provide feedback about the student's interpersonal and communication skills as well as his/her cardiovascular exam skills.

2. Watch each video in its entirety

- *We strongly recommend watching the video in full screen in a new browser tab*: Click the "YouTube" icon in the bottom right corner. After the video begins to play, click the frame icon in the bottom right corner to make it full screen.
- The standardized student's face is blurred in each video. We understand some information about the student's behavior is lost this way, but please do your best with the behavior you *are* able to observe.
- You will see 2-3 camera angles of the same encounter presented simultaneously

- Make sure to turn up the sound so you can hear what the standardized student and patient are saying
- You are not allowed to copy, retain, or distribute this video or video URL

3. After watching each the video, fill out the corresponding questionnaire
   - You can re-watch any part of the video you want while filling out each questionnaire
   - Please finalize the evaluation for each student before moving on to the next student. You will not be able to navigate back to previous evaluations.
4. After filling out evaluations for all of the videos, navigate to the final page to answer a few short questions about your experience as a standardized patient

If you cannot finish reviewing all of the videos in a single sitting and need to come back to finish at a later time, you can press the "Save and continue later" button found at the top of the page.

Pages 4-15: Interpersonal and Communications Skills and Physical Exam Skills Evaluations

[INTERVIEW AND PLAN VIDEO WITH ICSF] OR [PHYSICAL EXAM VIDEO WITH PESC]

Page 16: Demographics

[SP RATER DEMOGRAPHICS WAS HERE]

Page 17: Thank You

Appendix 12: Survey Gizmo Layout for the Physical Exam Skills Task for Crowd Raters

Page 1: Welcome + Screening

Have you, in the last 2 years, visited a US-based physician as a patient or as the guardian of a patient?
( ) Yes
( ) No

Page 2: Consent

[CONSENT WAS HERE]

Page 3: Physical Exam Skills Evaluation

General Instructions

You are about to watch a video of a medical student performing a portion of a physical examination (or physical exam, for short) on a patient actor. A patient actor is an actor who plays the role of a patient coming to see a doctor with a specific medical concern. Medical schools often hire patient actors to interact with medical students so that medical students can practice the skills they need to become good doctors (for example: communication, medical decision making, physical examination) in a "safe" environment before they interact with real patients. After each interaction, the patient actor provides feedback about the student's skills so the student can improve for the next interaction.

Because medical students and doctors are ultimately responsible to their patients, feedback from a patient's perspective about a medical student's skills are important as well. Therefore, we are interested in your judgment about the physical exam skills of the medical student in this video. A physical exam is a process by which a doctor investigates a patient's body for signs of disease.

1. Watch the entire Training Video and familiarize yourself with the Physical Exam Skills Questionnaire below. You will use this questionnaire to provide ratings of the student's physical exam skills.

- The Training Video is only intended to provide you with instruction on the correct way all of the physical exam maneuvers should be performed. Please note that the questionnaire is intended to be filled out based on the performance you observed in the Student Video, not in the Training Video.
- *We strongly recommend watching the videos in full screen in a new browser window*: Click the "YouTube" icon in the bottom right corner. After the video begins to play, click the frame icon in the bottom right corner to make it full screen.
- Make sure to turn up the sound so you can hear what's going on in the video
- For each of the maneuvers in the Student Video, you will make one of four choices on the questionnaire:
    - "Visible Done Correctly" - You saw clearly that the student performed this maneuver correctly

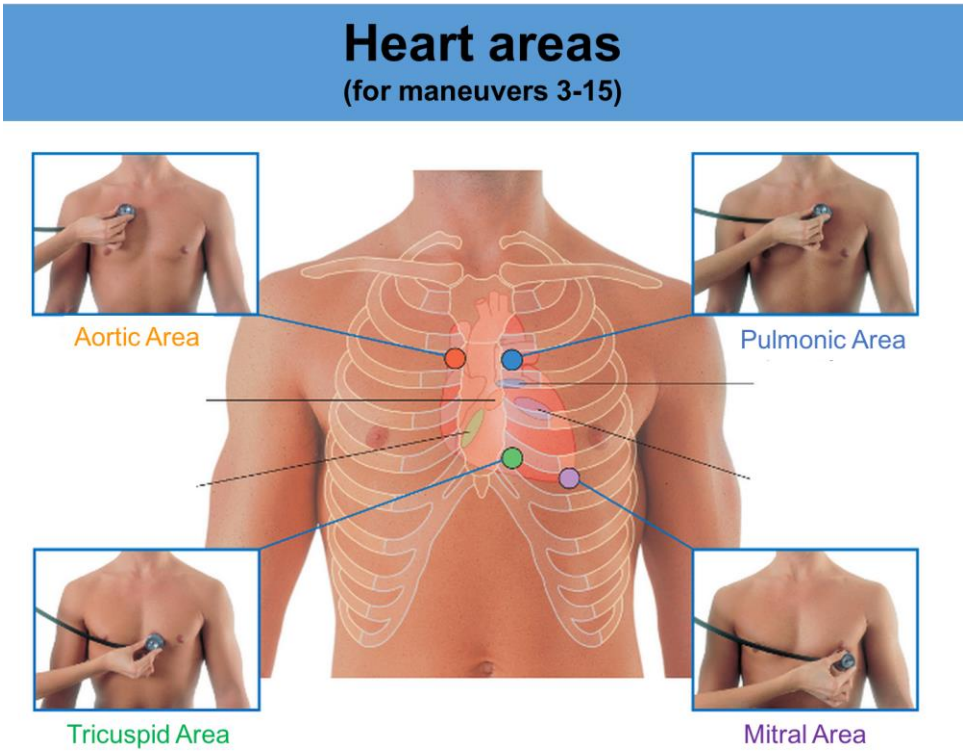
    - "Visible Done Incorrectly" = You saw clearly that the student performed this maneuver incorrectly
    - "Visible Not Done" = You saw clearly that the student did not perform this maneuver
    - "Observation Obscured" = The video production obscures a critical portion of this maneuver such that it is difficult for you to be sure whether the student performed the maneuver correctly or not
- For each maneuver scored as "Visible Done Incorrectly" or "Observation Obscured", please provide a brief account of what was done incorrectly or obscured (along with the item number of the maneuver).
- All physical exam manuevers must be performed on skin. For the purposes of this evaluation, a sports bra should be considered skin for female patient actors.
- Please review the section below titled, "Instructions Specific to this Physical Exam" for some additional information you will need to interpret the questionnaire.

2. Watch the entire Student Video
- The student's face is blurred to protect the student's privacy
- You are not allowed to copy, retain, or distribute this video or video URL
- You may see maneuvers in the Student Video that are not part of the questionnaire. Please ignore these.
- You will see 2-3 camera angles of the same encounter presented simultaneously

3. Fill out the questionnaire based on the performance you saw in the Student Video
- You can re-watch any part of the video you want while filling out the questionnaire
- Please be as accurate as possible. Students and faculty value your ratings. They rely on them to improve student performance and ensure that patients get the best care.

4. After filling out the questionnaire, navigate to the next page to answer a few short questions about your experience and some basic demographic information.
- You will not be able to navigate back to the video and questionnaire

Instructions Specific to this Physical Exam



Please answer the questions below to confirm you understand the instructions

	True	False
My goal is to accurately evaluate this student's physical exam skills	( )	( )
I should review the entire Training Video and questionnaire before watching the entire Student Video	( )	( )
I can view the videos in full screen in another browser tab by clicking the video's YouTube logo	( )	( )
I should distribute this video/video link to all of my family and friends	( )	( )
I should fill out the questionnaire on the Student Video, not the Training Video	( )	( )
The ulnar surface of the hand is on the same side as the thumb	( )	( )
All physical exam manuevers should be performed on skin	( )	( )
For female patient actors, a sports bra should be considered skin	( )	( )
The bell of the stethoscope is wider than the diaphragm of the stethoscope	( )	( )

When looking at the patient, the aortic area is in the bottom right area of the heart	( )	( )

Training Video

What kind of physical exam will be evaluated in this encounter?
( ) Abdominal Exam
( ) Cardiovascular Exam
( ) Extremities/Joint Exam
( ) Neurological Exam
( ) Respiratory Exam
( ) Vital Signs Exam

Physical Exam Video

[PESC WAS HERE]

Page 4: Rater Reactions

[RRX_PESC]

Page 5: Demographics

[CROWD RATER DEMOGRAPHICS WAS HERE]

Please feel free to leave general comments or feedback about this HIT here.

_____
_____
_____
_____

Page 6: Thank You

Appendix 13: Survey Gizmo Layout for the Interpersonal and Communication Skills Task for Crowd Raters

Page 1: Welcome + Screening

Have you, in the last 2 years, visited a US-based physician as a patient or as the guardian of a patient?
( ) Yes
( ) No

Page 2: Consent

[CONSENT WAS HERE]

Page 3: Interpersonal and Communication Skills Evaluation

Instructions

You are about to watch a video of a medical student interacting with a patient actor. A patient actor is an actor who plays the role of a patient coming to see a doctor with a specific medical concern. Medical schools often hire patient actors to interact with medical students so that medical students can practice the skills they need to become good doctors (for example: communication, medical decision making, physical examination) in a "safe" environment before they interact with real patients. After each interaction, the patient actor provides feedback about the student's skills so the student can improve for the next interaction.

Because medical students and doctors are ultimately responsible to their patients, feedback from a patient's perspective about a medical student's skills are important as well. Therefore, we are interested in your judgment about the social and communication skills of the medical student in this video.

1. Before watching the video, scroll down and familiarize yourself with the questionnaire you will use to provide feedback about the student's social and communication skills
    - You may find that the descriptions for each of the skill levels ("1-5") don't exactly describe the behavior of the student; that's OK. You should think of level "5" on the questionnaire as the best doctor you can imagine. Use the descriptions as a guide to help you make a decision about the student's general skill level. For example, if you feel that the student did a little better than the *types* of behaviors described in the "1" level but a little worse than the *types* of behaviors described in the "3" description, choose a "2".

2. Watch the whole video of the student interacting with the patient actor
    - You will watch the student gather information from and then give advice to the patient actor
    - While watching the video, try to put yourself in the patient actor's shoes. Think, "How would I feel about this interaction if I were visiting this student as a patient?"
    - The student's face is blurred to protect the student's privacy

- Students conduct a physical exam on the patient. The physical exam segment has been removed from this video, so you will see a brief cutaway/transition. This is normal.
- Make sure to turn up the sound so you can hear what the student and patient actor are saying
- You are not allowed to copy, retain, or distribute this video or video URL
- *We strongly recommend watching the video in full screen in a new browser tab*: Click the "YouTube" icon in the bottom right corner. After the video begins to play, click the frame icon in the bottom right corner to make it full screen.
- You will see 2-3 camera angles of the same encounter presented simultaneously

3. After watching the video, fill out the questionnaire
   - Put yourself if the patient actor's shoes. Think, "How would I feel about this interaction if I were visiting this student as a patient?"
   - You can re-watch any part of the video you want while filling out the questionnaire
   - Do the best you can, and be honest with your feedback. Students and faculty value your feedback. They rely on it to improve student performance and ensure that patients get the best care.

4. After filling out the questionnaire, navigate to the next page to answer a few short questions about your experience and some basic demographic information.
   - You will not be able to navigate back to the video and questionnaire

Please answer the questions below to confirm you understand the instructions

	True	**False**
My goal is to honestly evaluate this student's social and communication skills	( )	( )
I should read the entire questionnaire below before watching the entire video	( )	( )
The descriptions in the questionnaire are going to exactly match the student's behavior	( )	( )
I can view the video in full screen in another browser tab by clicking the video's YouTube logo	( )	( )
I should put myself in the patient's shoes as I fill out the questionnaire	( )	( )
I should distribute this video/video link to all of my family and friends	( )	( )

Interview and Plan Video

What's the main reason the patient came to visit the doctor's office today?
( ) Knee pain
( ) Chest pain
( ) Fatigue
( ) Headache
( ) Joint pain
( ) Back pain
( ) Persistent cough
( ) Stomach pain
( ) Trouble sleeping

According to the patient, what seems to make his/her condition worse?

_____
_____
_____
_____

Briefly summarize how the medical student/doctor recommend the patient treat his/her condition

_____
_____
_____
_____

[ICSF WAS HERE]

Page 4: Rater Reactions

[RRX_ICSF]

Page 5: Demographics

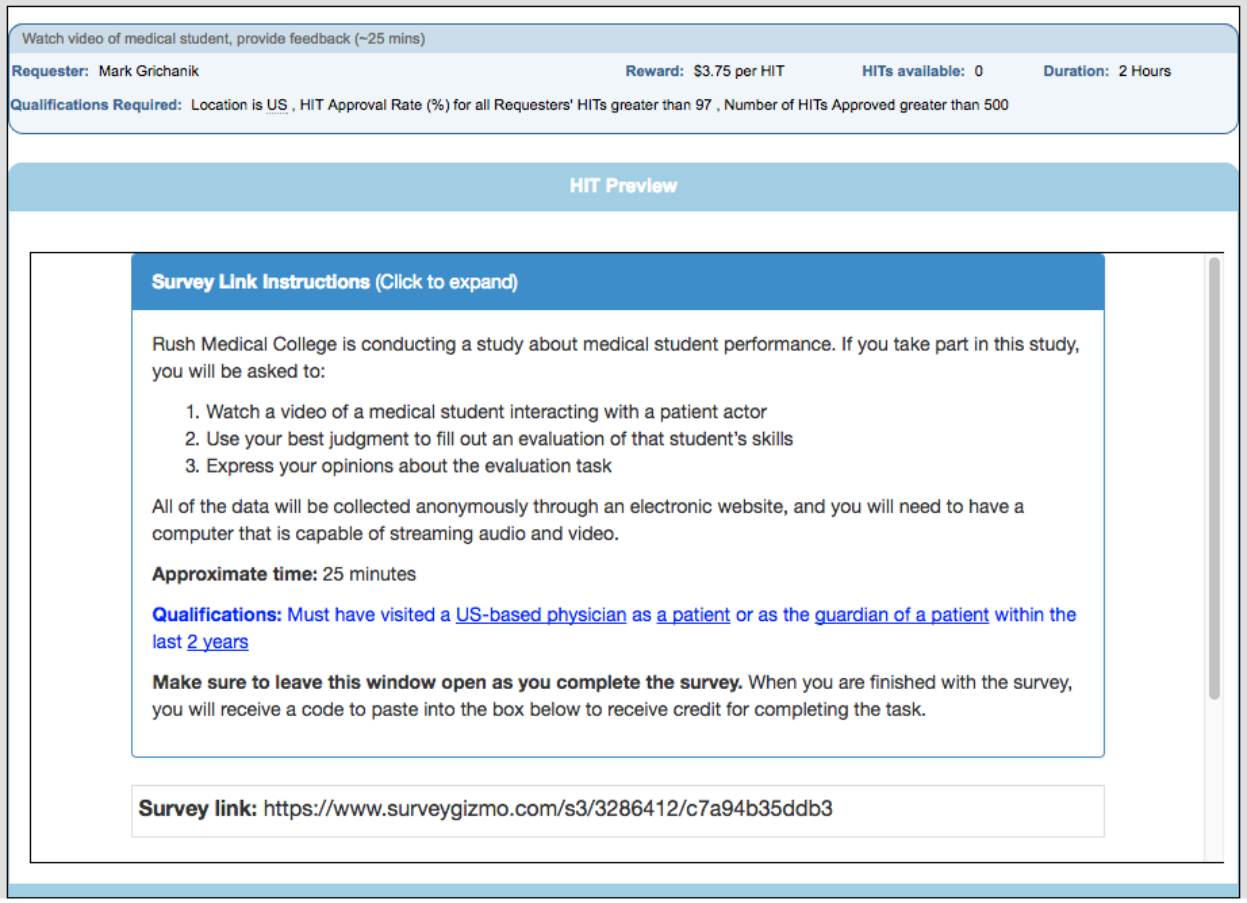
[CROWD RATER DEMOGRAPHICS WAS HERE]

Please feel free to leave general comments or feedback about this HIT here.

_____
_____
_____
_____

Page 6: Thank You

Appendix 14: Sample Amazon Mechanical Turk HIT Posting



Appendix 15: RRX-ICSF Content Analysis Comments by Theme

Apprehensive Evaluating

- I didn't think about being overly critical in evaluating since the student didn't seem to have much experience.
- I was a bit nervous in evaluating the student because of my lack of expertise.
- It's hard to know how much of any perceived lack of empathy / communication skills on the doctor's part is due to an actual lack vs. how much is due to the strange situation of interacting with a 'patient' who is known to be an actor. So, this was a little hard for me to judge accurately
- I understand the need to have this done as it is, with an actor but it is hard to truly evaluate a student doctor when they know it is not real. Would be interesting to see how different they would do if they didn't know they were being evaluated.
- At first I thought I may be a little uncomfortable evaluating someone else knowing that what I say reflects upon that individual.

Comfortable Evaluating

- I felt reasonably confident assessing the student's performance.
- I felt reasonably confident in assessing the social and communication skills of the student.
- I found it very easy to evaluate the student based on the questionnaire I received. I felt the questionnaire covered all of the important aspects of a skilled nurse or practitioner.
- I'm perfectly fine with evaluating a medical student
- I visit new doctors frequently for my own health problems, so I spend a lot of my time thinking about doctors bedside manners and if I like them and will continue a relationship with them, so this is very easy for me to make decisions quickly.
- I was comfortable with the assessment.
- I felt comfortable with the format and tried putting myself in the patient's place and using my own experience with my doctor as well.
- I felt mostly comfortable evaluating the medical student. If I did something like this more often I would probably be able to give better input.
- I felt mostly comfortable evaluating the student.
- I felt that I was rather easily able to assess the interaction depicted, and comment upon what I felt were the strengths and weaknesses of the student practitioner presented.
- was very comfortable adding my opinions to the survey. I
- I found evaluating the performance very easy

Enjoy

- enjoyed
- I liked the student and their interaction with the patient.
- Interesting hit, I enjoyed the task and hope more will be available in the future.
- I really really enjoyed this.
- I really enjoy these. Please post more!
- I liked this.
- It was a fun experience and I would definitely be interested in doing this again.

- This HITs are very interesting and I hope to see more of them in the future. Have a great day/night!
- it was interesting and it paid well
- Thanks for a really fairly paid and interesting survey! I very much enjoyed it.
- I liked evaluating the student.
- I'm having fun watching the interactions between the patient actor and the student, especially after having already done one and being able to compare the different styles in the students.
- I really enjoyed this task. ***Thanks again.
- Enjoyed it. ***Great!
- I really like this type of surveys, I feel like I am doing something useful :)
- This is a well made survey, I really enjoyed it. Thank you!
- ...and overall an easy and fun evaluation experience.
- THIS WAS A VERY ENJOYABLE TASK AND I HAD A GREAT TIME DOING THIS HIT.
- THE TASK WAS VERY, VERY INTERESTING AND WITHOUT A DOUBT MY FAVORITE HIT I HAVE EVER DONE ON MTURK.
- Thank you. I enjoyed the task.
- I really enjoyed taking part in this HIT. It was very unique compared to the typical studies I take part in. Thank you!
- I thought this was fun and I hope I helped you with your research
- I enjoyed this and I thought the instructions were great :)
- I liked this opportunity to review future doctor's communication skills.
- I liked evaluating her communication as a potential patient. I liked that it gave me a preview of how the doctor may communicate and the tool itself may be useful for other new patients to evaluate their new doctor.
- I liked this.
- Thanks, I liked this a lot.
- Thank you for the opportunity to participate!
- I actually enjoyed watching the interaction and am happy to have given my feedback
- The HIT was fun to complete. Thanks!
- I really enjoy doing these HITs. Thanks!
- This was a very positive and enjoyable experience for a myriad of reasons.
- Very interesting, enjoyable, engaging, and well-designed task.
- it is a task I somewhat enjoy doing. :)
- It was a very satisfactory task on a personal level. I was very happy on helping to provide feedback for future doctors
- The study was very fun
- .I was happy to participate on it
- it was fun and original
- Great study. Thank you.
- This was a fun hit and I would like to do more in the future.
- Fun hit. Thanks!
- Good study.

Face Blurring

- The blurring of the student's face made assessing the amount of eye contact somewhat difficult, although much of the time the face was obviously on the check sheet and not facing the patient.
- Although the student doctor's face was blurred, I tried to pay attention to body language.
- It would have been helpful to be able to see the student's facial expressions.
- The blurred face made assessing eye contact difficult
- The image blurring of the student's face makes it a little difficult to judge their social skills in the eye contact and expressions area.

Interesting

- Interesting and enjoyable HIT
- Interesting hit, I enjoyed the task and hope more will be available in the future.
- It was interesting
- It was very interesting and well-done. Thanks!
- This was a very interesting HIT. Thank you!
- This is a very interesting assignment.
- THE TASK WAS VERY, VERY INTERESTING AND WITHOUT A DOUBT MY FAVORITE HIT I HAVE EVER DONE ON MTURK.
- VERY INTERESTING TASK!
- I found this to be quite an interesting experience.
- Interesting and empowering.
- Interesting and empowering. I like being able to influence the health care system, even if in a very small way.
- Thank you. I found the task to be quite interesting and involving.
- well compensated and interesting to complete
- It was interesting and felt like I was being useful.
- I found this extremely interesting.
- really interesting! I hope my review is helpful!
- This was a very interesting and engaging experience. Additionally, this is a very interesting initiative and one that should be explored given its benefits to the development of better physicians.
- Very interesting, enjoyable, engaging, and well-designed task.
- it was original
- I've done similar work, but this was definitely more in-depth and real. It is interesting and a little worrisome to see the lack of communication skills of some people.

Objective

- I tried to be as objective as possible and put myself in the 'patient's shoes,' but it was also difficult to not feel some understanding toward the medical school student as well and her nerves and simply learning an incredibly difficult job with tons of responsibility.
- Many elements of the task seemed fairly subjective, but I tried to give feedback in an open and honest way.

- I am glad that I could participate and give an honest opinion regarding the medical student's social and communication skills
- I felt as though I was being fairly critical but I scored as honestly as I could. I have an excellent physician and that was my only yardstick.

Patient Voice

- In terms of the task itself, I think this is a great idea. Most people only visit the doctor whenever something could be wrong, and it's very very important that the patient feels confidence and friendliness from the physician. My grandma has been in and out of the hospital over the last several months, and while most of her doctors have been friendly, there has been a few that just feels like you're the means of a paycheck.
- I think it is a great idea to evaluate students like that.
- made me think of my own doctor and questions I should ask as a patient.
- I feel it is good for medical students to be evaluated on their performance so they can build empathy and interpersonal communication skills with their patients. Patients are very vulnerable when they come to doctors with concerns on their health and the transition from diagnosis to treatment can go smoother when the patient and doctor understand one another.
- I've had some TERRIBLE meetings that had to do with this exact problem, so I know how much pain is involved with GERD and how scary it is before you get an answer or know how to fix it.
- I believe that it is very necessary for future doctors as a lot do not have a good sense of interaction with their patients.
- I think my feedback is valuable as I've had a lot of issues with the demeanor and attitude of a lot of doctors and nurses. It makes me happy to know that people out there actually care about the patients point of view.
- I think this evaluation tool is great and totally necessary.
- I hope these kinds of evaluations continue to be done as they provide the public with a great service in information and training.
- Its an interesting event to watch as a student learns to become a doctor. I never really think of the doctors as being the nervous ones until now. Hopefully with enough surveys like these and more training, doctors can begin to show a more caring side to themselves.
- I like this way of evaluating medical professionals. This is a very useful method and can do a lot of good.
- I think this is a good way to evaluate medical students.
- As I watched the video and put myself in the patient's shoes it became easier and I looked at it as a teaching tool in a way. Something the student could use to work on things that may be a little lacking. We all have things we can work on and this is a good way of finding strengths and weaknesses.
- Overall I think this is a good tool for evaluating students.
- I think having these evaluations is a great idea. Hopefully we are allowed to do more of them. Thank you!
- I felt the evaluation tool was a great way to measure the students skill level.
- My experience was positive, I don't have much to say about it. I'm happy to be able to contribute to this research cuz it seems important.

- I liked this opportunity to review future doctor's communication skills. 'Bedside manner' is an important skill for doctors to develop aside from technical knowledge.
- I liked evaluating her communication as a potential patient. I liked that it gave me a preview of how the doctor may communicate and the tool itself may be useful for other new patients to evaluate their new doctor.
- I liked being able to provide some feedback on the relationship aspect of practice for doctors. Communication is just as important as the technical expertise. I think this tool will be helpful in improving the interpersonal practices of doctors.
- This tool is a very good way to give patients a way to evaluate the communication skills of future doctors. Relationships are important to building trust, doctors need this skill to keep patients healthym
- It was interesting and felt like I was being useful.
- I visit new doctors frequently for my own health problems, so I spend a lot of my time thinking about doctors bedside manners and if I like them and will continue a relationship with them, so this is very easy for me to make decisions quickly.
- I feel like the general population should be involved in giving feedback to medical students because it can only help to improve their skill set. This is definitely something that I would be happy to do again.
- I actually think this would be a great asset in the real world.
- I actually enjoyed watching the interaction and am happy to have given my feedback
- Additionally, this is a very interesting initiative and one that should be explored given its benefits to the development of better physicians.
- It was a very satisfactory task on a personal level. I was very happy on helping to provide feedback for future doctors
- Working on this task was a very satisfactory personal experience for me.I was really happy that i could help future doctors

Student Performance

- I found the question about quitting smoking to be a bit too direct and that her primary issue should have been handled first before mentioning it.
- The encounter was great however the conclusion of the visit with treatment options could be improved upon. The student did seem genuinely concerned for the patient.
- I would be surprised if the patient would give the encounter a positive score, I felt a lack of empathy and an absence of care.
- In my opinion this student did a great job. I understand the summary is important but maybe make it a smoother transition
- I would love to be treated this way by my doctor!
- i thought he did a professional job ,well spoken, friendly
- I also meant to say that the student at times sounded too much like he was speaking to a child...like 'dripping' with politeness.
- This was my third evaluation and this student did the best by far out of the ones I saw.
- The student was very professional. She can work on being more informative and connecting with the patient however.
- They were very pleasant

- It really just seems to me this student is young and not really experienced. Possibly nervous herself.

Task Negative

- I did another one of these earlier, and am still not 100% sure about the summary, recap, exit question block, but that'd be my only complaint.
- The only thing I wasn't completely sure about was the 'Ending the Encounter' questions. The student did some of those things during the patient evaluation, but they didn't happen at the end of evaluation so I wasn't sure if I should answer yes because it had happened earlier or no since it didn't happen at the end.
- I wasn't sure whether the question about whether the student provided instructions for what the patient should do when leaving the room meant instructions about the treatment action plan or logistical instructions. Based on context, I assumed logistical instructions and answered
- The experience was fine. The layout of the questions on the first page might be a bit too spaced together which made it a bit harder to read at first.
- I found it somewhat odd that you ask participants to only provide 100 characters to describe the patient's condition, but allow 100 words when describing treatment conditions. That's quite a disparity.
- The layout of the questionnaire was a little overwhelming. I'm not sure the examples were entirely necessary, maybe the categories should just be more concise
- The evaulation could be formatted in a way that streamlines the survey experience. Maybe give a short test first to see if the survey taker can comprehend what they are doing and to familiarize themselves with the questions before viewing the video.
- %It was very difficult to come up with a positive thing to say. I wish that box either wasn't mandatory or there was a box that said 'there was nothing positive to comment on'.

Task Positive

- The instructions were clear however took a bit to get used to. Now that I am familiar with the type of encounter, I will be able to understand it much faster.
- The tools all worked perfectly
- I felt that the tools, the instructions, etc were all very good.
- It was easy to understand, and well laid out.
- It was pretty straightforward.
- It was easy to understand the instructions, the questions were well laid out, and if approved it pays well.
- I think the options provided in the bubble scale questions were sufficient to evaluate the student doctor.
- Having all questions and the video on one page made it easy to check my answers and review the video. The instructions were clear and easy to understand.
- The instructions were helpful. I found that the questionnaire helped me to zone in on what areas to focus on when evaluating the student doctor.
- The evaluation was a fine experience.
- Adequate instructions
- Very adequate instructions

- Instructions were very adequate
- Very adequate instructions and everything is laid out in a very intuitive manner for us.
- The survey was very straight-forward and easy to understand.
- Great HIT. Thanks!
- Everything went very smoothly for me, mostly because you let me know to read ahead and get an idea of what to look/listen for in the video, felt a lot like studying for a test where you already know the questions, which obviously makes it a very easy experience. I had no technical or moral issues judging the student either, so everything went very well.
- HIT, set up nicely and everything was explained really clearly.
- The instructions were great, and they let me know exactly what to expect and what my job was.
- Really nice HIT, everything made sense and went along quickly.
- The experience was fine, no problems.
- The instructions were easy to follow and understand.
- Everything seemed to go fine.
- The experience was simple and easy to understand. Other than not being sure what to say about how the student could improve because I honestly thought that he did such an excellent job, I felt that the questions were easy to answer and that the questionnaire was laid out well.
- Great study! Very well designed.
- The experience was okay. There were no issues.
- I had no issues with the experience.
- There were no issues.
- I thought this questionnaire was well-formatted and does a good job of telling participants what is expected of them when using the tool.
- I felt this was very straight forward, the instructions were clear
- Well this is my third one, so I have it down :)
- It all made sense, I think I have the hang of it
- Fairly straight forward
- All good.
- I thought the instructions were well thought out and clear.
- Instructions were clear. The questions were easily worded and guided through the process (in sequence)
- THE INSTRUCTIONS FOR THIS TASK WERE VERY CLEAR AND I HAD NO ISSUES UNDERSTANDING THEM.
- Everything was pretty straight forward. The directions were clear, the rating scales were clear, and I knew exactly what was expected of me
- Everything was clear and straight forward.
- I thought the evaluation questions were extremely clear. The rating choices were written so that it was simple to understand the specific differences between each rating.
- The instructions given to complete this evaluation were very clear.
- I enjoyed this and I thought the instructions were great :)
- The questionnaire was easy to follow, I found each question relating to specific events in the medical examination.

- The questionnaire was very informative and I was able to find an event in the video for each question.
- I think the instructions were clear.
- It was very simple and straightforward.
- Very simple and clear
- Instructions were clear
- The instructions were clear and the evaluation tool were easy to use.
- The instructions were clear and the evaluation tool was easy to use.
- The instructions were very clear and easy to understand.
- Overall it was very easy to use the evaluation tool.
- I felt that I was rather easily able to assess the interaction depicted, and comment upon what I felt were the strengths and weaknesses of the student practitioner presented.
- Everything was very straightforward and easy to understand.
- The HIT was very easy to follow.
- In addition, the instructions, OSCE evaluation questionnaire, and etc. were clear.
- Very interesting, enjoyable, engaging, and well-designed task.
- All was understood
- the tasks was simple.
- easy to do
- I found all instruction to be very clear. I found the study to be very straight forward
- I felt the video gave me a good amount of information in order to make a good evaluation. I understood my task clearly.
- The instructions were straight-forward. I knew what I was supposed to do. The video worked fine and the questions were related to the video.
- Good study.
- No comments. It was very well explained and easy to do.
- Everything was clear and easy to understand.
- The evaluation is well done and formatted. All the instructions are clear and easy to understand.
- I feel that everything presented was very clear and easy to understand.
- There were no problems. The instructions were all clear! Everything went smoothly.
- It went well! Nothing was unclear!

Technical Concern

- While reviewing medical type videos, I find it beneficial to have the video in the left frame, with the questionnaire on the right. It makes scoring and reviewing the video much easier. A good example of this is the type of surgical evaluations from C-SATS, which allow instant review and then scoring on the right. The video will keep in frame while the questions can be scrolled on the right.
- I will say that I was hesitant to put the video in full screen(but I am working on a very large monitor anyway). I wanted to be able to answer questions as I heard the information, so that I would not forget anything. I ended up opening it separately into a different monitor screen so I could have both the video and the questionnaire up at the same time.

- The only small issue encountered was not related to the evaluation tools per se, but rather a strange issue related to Youtube which I was able to troubleshoot and resolve.

Appendix 16: RRX-PE Content Analysis Comments by Theme

Apprehensive Evaluating

- I felt like I knew what was expected, but then actually evaluating someone, I felt like I had no medical knowledge. A little intimidating.
- I felt nervous for the student since he was being evaluated on camera but thought he did a good job.
- I still feel unsure about my ability, but I am trying to be thorough.
- Everything went well I think, although I felt almost like I didn't know enough to know exactly what was being done at all times.
- Obviously I can't make specific determinations on medical skills and abilities but it is easy enough to watch a training video and then see if those some procedures are being applied at least on a surface level. So there is some way to evaluate, but mostly in a broad sense.
- The only issue I had with this tool is that the training video was not entirely explicit about whether or not conducting a cardiovascular exam using the patient's back was an acceptable practice or not. In my own physical examinations, I have had a doctor use both sides when conducting a similar examination. The medical student only used the patient's back, making it difficult to provide an accurate evaluation of their method.
- I was nervous as I didn't want to rate the doctor bad or miss anything
- I just am unsure judging another.
- I feel like I didn't really know what I was doing.
- It made me nervous a bit just because I wanted to make sure I was judging it completely accurate to help the student better understand what they need to improve on if anything.
- I felt like I had to find something wrong. Why would the student make that egregious of a mistake? I suppose they could have done it incorrectly for training purposes.

Comfortable Evaluating

- I watched carefully for any mistakes so I feel I did a good job.
- I think I am getting a little more confident at evaluating the students.
- I felt confident evaluating the students performance
- I found them to perfectly fine for understanding what I needed to do.
- The combination of the training video and the images made me feel very confident with what I was looking for.
- I felt comfortable evaluating the student
- Looked fairly easy to judge.
- I felt like I learned enough from the instructional video in order to accurately evaluate the medical student's performance.
- I am a nurse and felt very confident while watching the video and providing feedback regarding the student video.
- This is my second evaluation, and I did feel more comfortable and confident in my assessments this time around
- I have medical experience as well so it made it easier for me to evaluate the student.

Enjoy/Fun

- fun study
- I thought this was a good HIT and paid well. I hope that you will post more HITs like this in the future.
- I enjoyed the experience
- This was a fun task.
- I enjoyed evaluating what happened.
- This was a great study and I really enjoyed it. I wouldn't mind doing more of these evaluations.Great hit.
- I enjoyed this evaluation.
- Thanks for the opportunity to participate; good luck with your research!
- I like these surveys.
- This was nice a i learned a lot
- I did enjoy this exercise
- Fun thanks - cheers
- I enjoyed watching the videos.
- I learned a lot and enjoyed it also.
- I think this was great
- I enjoyed watching the student's performance and checking it against the list of what they had to do.
- Thank you for this great HIT. I would love to be involved in more of these in the future. Please feel free to reach out. Many thanks!
- I enjoyed this survey
- This was my first such HIT and I hope to do more!
- I enjoyed the experience.
- this was a fun hit.

Interesting

- Very interesting study.
- It's an interesting task.
- Interesting task.
- I am hoping I did it correctly.
- These HITs are very interesting and I hope there will be more in the future! Have a great day!
- I found the exam very interesting concerning the process.
- This was very interesting.
- This was very interesting
- This was a really interesting task.
- This was interesting
- The hit was interesting
- I found this interesting
- This was very interesting
- Very interesting!
- This was a very interesting task.
- Thanks! This was a very interesting task!

- This was very interesting.
- This was very interesting. Thanks!
- it was different and interesting.
- I found this really interesting.
- It was very interesting to learn different things.
- It was interesting to see the doctor in training.
- I thought that it was very interesting
- It was neat to watch this common procedure
- Very cool task

Objective

- I wanted to be fair
- I hope I was fair.

Patient Voice

- I did not know a regular joe like myself could rate these sessions accurately but now I feel like I can!
- It made me feel like patients should get the opportunity to do as such more often.
- I think it's good for patients to be able to evaluate students, as well as doctors.
- I hope that my contribution will be os some value.
- &I think evaluating medical students on their performance can be helpful to them. Having someone watch their examinations will give them feedback to learn from especially if they did something wrong or could have done something better.
- I feel that evaluating medical students physical exams skills will provide them with constructive criticism to help them to become better at their examining skills.
- I love to help out
- I am happy to help the students learn from their mistakes and hope that my evaluations help them become the best doctors they can be
- a good way to evaluate new doctors.

Student Performance

- proper methods throughout
- student was too quick, didn't do everything
- The very first of the video shows that last moments of the student rubbing her hands together after applying hand sanitizer. It would be more clear to evaluators if the video showed the applying of sanitizer so we can be certain.
- student did not perform most procedures
- The student did a fantastic job though seemed a little nervous.
- I could feel that the student was sort of nervous but she did everything correctly.
- the student did a good job
- I think the student who performed the physical exam in this video did a good job. She performed really well in the exam.
- Student was out of order but got most steps completed eventually.
- I felt that this student didn't quite know where to put the stethoscope to get the procedures right and he seemed to forget some steps.

- The student did well. She however did put the patient in an uncomfortable position when she had the patient lay down and then untie the gown. The gown should have been untied when the patient was sitting.
- I feel like the student did a decent job, just left out certain maneuvers that were done in the training video, so I evaluated based on that.
- The student performed all of the tasks correctly.
- This doctor seemed to rush through the process.

Task Negative

- I cannot understand what I could not find any errors
- I did not see in the training video or the student video the part where they use the palm of the hand ulnar to inspect the patient, but it was asked in the questions. I was unsure of that. I saw her use the fingertips and the instruction video used finger tips as well. Maybe I misunderstood that aspect.
- The anatomical diagram would be better if it were closer to the questions so the reviewer can refer to it more easily.
- I wish the heart diagram image was closer to video for better consulting, but otherwise it is good.
- would like to do more with more instruction though this video was good.
- I was not too sure what the questionnaire meant by 'finger pads'.
- I did notice the student did not listen to the heart in the same order as the training video did. I don't know if that matters at all or not though.
- Should there be a comment box to explain the steps that the student omitted such as in this video?

Task Positive

- I thought the training video was well done
- Very clear
- Everything went well.
- I thought the training video was helpful to get a good idea about how the student should perform the exam. The diagrams were also helpful.
- instructions were complete, made it easy to evaluate
- The training video was very helpful in determining proper processes for the physical exam.
- The instructions were very clear and helpful.
- The instructions were long, but clear. The diagrams helped and the videos were both clear and easy to understand. The questions were straight-forward and easy to answer.
- All was good.
- Clear instructions given.
- All instructions were clear.
- I felt the information and diagrams were adequate for performing the evaluation. The training video was clear and easy to follow.
- The combination of the training video and the images made me feel very confident with what I was looking for.

- The instructions looked overwhelming at first glance, but weren't nearly as complicated as I expected them to be. And I appreciated that they were kept on the same page as the training video and the student video so that I could easily reference them when needed.
- There were no issues evaluating the student
- It was very weel explained, and a great HIT
- I felt that all instructions were clear, and the training video was clear as helped judge the student video.
- I found the instructions to be clear and easy to follow, and the tool was very easy to use.
- This experience was fine. I was given plenty of material to evaluate the students properly.
- was a fine experience. I was given enough material to accurately review the students.
- The experience evaluating this encounter was fine. I was given plenty of material to get a feel for how the process should be done correctly.
- experience was fine. I was given plenty of instruction on how to evaluate the medical students properly
- The experience for evaluating this encounter was fine. The training material at the start was perfect.
- I felt comfortable evaluating the student with the information provided to me at the beginning of the study.
- the instructions were clear
- everything was fine.
- Seemed straightforward.
- Seemed straightforward
- Seemed straight forward.
- Seemed pretty straight forward.
- Seemed straightforward.
- Everything was detailed well. It helped that the doctor in training did everything by the book. I had to pause at some points to get a good look on which side of the stethoscope was being used. It worked out.
- Instructions were clear and everything went smooth.
- All instructions were clear and easy to follow.
- Instructions were clear.
- Instructions provided were clear and allowed me to accurately evaluate the students performance.
- Everything went smooth and the instructions were clear
- Instructions were very clear and easy to understand. Training video was extremely helpful in preparation for the evaluation.
- The tool for eval is good. The training video was good.
- Instructions were good.
- good Hit, no problems
- I felt like I learned enough from the instructional video in order to accurately evaluate the medical student's performance.
- The video was clearly visible and the 3 different views really helped make the evaluation accurately.
- I thought that the instructions were lengthy, but necessary to be clear.

- I think this was a great hit with very clear instructions, both written and with the training video.
- Instructions were clear.
- The instructions were very clear.
- I felt like the instructions were very clear.
- The directions were very clear. Thank you.
- Easy to understand. Helpful training video
- Everything was easy to understand. The training video was very helpful.
- Clear instructions. Helpful training video
- I appreciate the clear instructions.
- The instructions were well written and allowed me to read and understand them in an efficient manner.
- I felt like the instructions were clear and easy to follow.
- The instructions and specially the training video is very helpful.
- The training video was very helpful in understanding exactly what I was looking for.
- I thought everything was well explained.
- Learning the training was interesting

Technical Concerns

- Everything was great except the hand washing question. It's not really fair because we don't know exactly what point in the exam the video starts.

Appendix 17: Rating and Feedback Quality Questionnaire for the ICS Task (RFQQ-ICS)

Please indicate the degree to which you agree with each of the following statements for each feedback package.

This feedback package features <u>numerical ratings</u> that <u>reflect the performance</u> you saw in the video.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Package A	( )	( )	( )	( )	( )	( )
Package B	( )	( )	( )	( )	( )	( )
Package C	( )	( )	( )	( )	( )	( )

This feedback package features <u>written feedback</u> that <u>reflects the performance</u> you saw in the video.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Package A	( )	( )	( )	( )	( )	( )
Package B	( )	( )	( )	( )	( )	( )
Package C	( )	( )	( )	( )	( )	( )

This feedback package features written feedback that is diverse.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Package A	( )	( )	( )	( )	( )	( )
Package B	( )	( )	( )	( )	( )	( )
Package C	( )	( )	( )	( )	( )	( )

This feedback package features written feedback that is high quality.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Package A	( )	( )	( )	( )	( )	( )
Package B	( )	( )	( )	( )	( )	( )
Package C	( )	( )	( )	( )	( )	( )

This feedback package features written feedback that is specific.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Package A	( )	( )	( )	( )	( )	( )
Package B	( )	( )	( )	( )	( )	( )
Package C	( )	( )	( )	( )	( )	( )

This feedback package features written feedback that would be useful in guiding the student in the video to change his/her performance in the future.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Package A	( )	( )	( )	( )	( )	( )
Package B	( )	( )	( )	( )	( )	( )
Package C	( )	( )	( )	( )	( )	( )

Imagine you were the student in this video and wanted feedback about your performance. If you could receive only one of these feedback packages, which feedback package would you prefer?
( ) Package A
( ) Package B
( ) Package C

Why do you prefer the feedback package you chose above?
(Feel free provide your impressions about any of the individual feedback packages or about any differences that stood out to you among the feedback packages. Please refer to the feedback packages by name (e.g., "Package A was...")
_____
_____

Appendix 18: Rating and Feedback Quality Questionnaire for the PE Task (RFQQ-PE)

Type in the number of evaluators/raters that contributed to each feedback package:
Package A: _____
Package B: _____
Package C: _____

Please indicate the degree to which you agree with each of the following statements for each feedback package.

This feedback package features numerical ratings that reflect the performance you saw in the video.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Package A	( )	( )	( )	( )	( )	( )
Package B	( )	( )	( )	( )	( )	( )
Package C	( )	( )	( )	( )	( )	( )

Imagine you were the student in this video and wanted feedback about your performance. If you could receive only one of these feedback packages, which feedback package would you prefer?
( ) Package A
( ) Package B
( ) Package C

Why do you prefer the feedback package you chose above?
(Feel free provide your impressions about any of the individual feedback packages or about any differences that stood out to you among the feedback packages. Please refer to the feedback packages by name (e.g., "Package A was...")
_____
_____

Appendix 19: Student Reactions Questionnaire (SRX)

Please indicate the degree to which you agree with each of the following statements:

Part 1	**Strongly Disagree**	**Disagree**	**Slightly Disagree**	**Slightly Agree**	**Agree**	**Strongly Agree**
SP ratings and feedback would be **valuable**, even if I had ratings and feedback from the **crowd**. SRX_1	( )	( )	( )	( )	( )	( )
Faculty ratings and feedback would be **valuable**, even if I had ratings and feedback from the **crowd**. SRX_2	( )	( )	( )	( )	( )	( )
Crowd ratings would be **valuable**, even if I had ratings and feedback from **SPs**. SRX_3	( )	( )	( )	( )	( )	( )
Crowd ratings would be **valuable**, even if I had ratings and feedback from **faculty**. SRX_4	( )	( )	( )	( )	( )	( )
Crowdsourced ratings and feedback are **worthwhile**. SRX_5	( )	( )	( )	( )	( )	( )
Generally, I receive a **sufficient amount** of **ratings and feedback** about my clinical skills as part of simulated patient encounters. SRX_6	( )	( )	( )	( )	( )	( )
The **ratings and feedback** I generally receive as part of simulated patient encounters give me the information I need to **help me improve my performance**. SRX_7	( )	( )	( )	( )	( )	( )
The ratings and feedback I receive from **SPs** as part of simulated patient encounters are typically an **accurate representation of my performance**.	( )	( )	( )	( )	( )	( )

SRX_8						
I am willing to receive **negative feedback** from **SPs**, even if the comments might upset me. SRX_9	( )	( )	( )	( )	( )	( )
I am willing to receive **negative feedback** from **faculty**, even if the comments might upset me. SRX_10	( )	( )	( )	( )	( )	( )
I am willing to receive **negative feedback** from **crowd** raters, even if the comments might upset me. SRX_11	( )	( )	( )	( )	( )	( )
It is important that I get **feedback** from **multiple sources** about my clinical skills. SRX_12	( )	( )	( )	( )	( )	( )
The **patient's perspective** about my clinical skills is **important**. SRX_13	( )	( )	( )	( )	( )	( )
I trust that **SP** raters are **fair and objective** when providing ratings and feedback. SRX_14	( )	( )	( )	( )	( )	( )
I trust that **faculty** raters are **fair and objective** when providing ratings and feedback. SRX_15	( )	( )	( )	( )	( )	( )
I trust that **crowd** raters would be **fair and objective** when providing ratings and feedback. SRX_16	( )	( )	( )	( )	( )	( )

Part 2	**Strongly Disagree**	**Disagree**	**Slightly Disagree**	**Slightly Agree**	**Agree**	**Strongly Agree**

Crowdsourced ratings and feedback would be acceptable for **formative assessment** purposes (i.e., for practice/feedback) SRX_17	⊃	⊃	⊃	⊃	⊃	⊃
Crowdsourced ratings and feedback would be acceptable for **summative assessment** purposes (i.e., scores count toward final grade/promotions decisions) SRX_18	⊃	⊃	⊃	⊃	⊃	⊃
I feel that **crowd** raters are **appropriate** judges of **interpersonal and communication skills**. SRX_19	⊃	⊃	⊃	⊃	⊃	⊃
I would feel **comfortable being evaluated** by a **crowd** rater on my **interpersonal and communication skills**. SRX_20	⊃	⊃	⊃	⊃	⊃	⊃
I feel that **crowd** raters are **appropriate** judges of **physical examination skills** SRX_21	⊃	⊃	⊃	⊃	⊃	⊃
I would feel **comfortable being evaluated** by a **crowd** rater on my **physical examination skills**. SRX_22	⊃	⊃	⊃	⊃	⊃	⊃
It is **important** for me to **receive ratings and feedback** in a **timely manner**. SRX_23	⊃	⊃	⊃	⊃	⊃	⊃
Generally, I **receive ratings and feedback** from my simulated patient encounters in a **timely manner**. SRX_24	⊃	⊃	⊃	⊃	⊃	⊃

I would be **concerned about privacy** issues if my simulated clinical encounter performance video was evaluated by **crowd** raters with my **face visible**. SRX_25	○	○	○	○	○	○
I would be **concerned about privacy** issues if my simulated clinical encounter performance video was evaluated by **crowd** raters with my **face blurred**. SRX_26	○	○	○	○	○	○
The **loss of information** (e.g., about eye contact) a **crowd** rater experiences when my **face is blurred** is a **worthy tradeoff** to **protect my privacy**. SRX_27	○	○	○	○	○	○
I believe **SP** raters sometimes **adjust performance ratings and feedback** up or down due to **social forces** (e.g., avoiding negative interactions with students). SRX_28	○	○	○	○	○	○
I believe **faculty** raters sometimes **adjust performance ratings and feedback** up or down due to **social forces** (e.g., avoiding negative interactions with students). SRX_29	○	○	○	○	○	○
I am **motivated** to use the **SP** ratings and feedback from to **improve my performance** SRX_30	○	○	○	○	○	○
I am **motivated** to use the **faculty** ratings and feedback from to **improve my performance** SRX_31	○	○	○	○	○	○
I would be **motivated** to use the **crowdsourced** ratings and feedback from to **improve my performance** SRX_32	○	○	○	○	○	○

What additional advantages can you think of for using a crowdsourced system for reviewing standardized patient encounters?

_____
_____
_____
_____

What disadvantages can you think of for using a crowdsourced system for reviewing standardized patient encounters?

_____
_____
_____
_____

Appendix 20: Sample Interpersonal and Communication Skills Feedback Package Presentation

Item	Your Ratings	Package A	Package B	Package C
Number of evaluators	1 (You)	1	1	18
Beginning the encounter: The student addressed the patient respectfully using the patient's last name	Student's Rating	Yes	Yes	Yes (94%) No (6%)
Beginning the encounter: The student introduced him/herself by name	Student's Rating	Yes	Yes	Yes (100%) No (0%)
Beginning the encounter: The student described his/her role on the healthcare team	Student's Rating	Yes	Yes	Yes (100%) No (0%)
Gathering Information About Patient's Concerns	Student's Rating	3	3	4.3 (0.8)
Empathy for the Patient's Distress	Student's Rating	3	3	4.2 (0.7)
Providing Patient with Information	Student's Rating	3	2	4.2 (0.7)
Using Appropriate & Sensitive Language	Student's Rating	4	4	4.1 (1.1)
Making Decisions	Student's Rating	3	3	3.8 (0.8)
Supporting Emotions & Fostering Relationship	Student's Rating	3	4	4.4 (0.6)
Ending the Encounter: The student summarized the patient's reason for visit	Student's Rating	Yes	No	Yes (100%) No (0%)
Ending the Encounter: The student checked for accuracy of their	Student's Rating	Yes	No	Yes (100%)

summary				No (0%)
Ending the Encounter: The student provided the patient with instructions on what to do after he/she left the room	Student's Rating	No	Yes	Yes (83%) No (17%)
Overall Interpersonal and Communication Skills	Student's Rating	3	4	3.9 (0.9)
Probability of Return	Student's Rating	3	3	4.1 (1.0)

With respect to interpersonal and communication skills, <u>what did this student DO WELL</u> in this patient encounter?

Your Feedback:

- Student's Feedback

Package A:

- The student seemed genuinely empathetic, as when she talked about "getting to the bottom" of the patient's problem, or when she said she was "sorry to hear" of the patient's nausea. I also appreciated that when she summarized what the patient had told her so far, she said, "please let me know if I've got anything wrong."

Package B:

- Pleasant and conversational. Brisk pace did not feel off-putting.

Package C:

- Accurately listened to the patient's concerns and showed interest in the patient's well-being and care. Body language was pleasant and their attitude was very cordial.
- Was a great listener. Very conversational.
- She introduced herself and asked if she could take some notes. She did say things like "oh, good" when the patient said she wasn't currently in pain, which is nice. I thought it was good she asked if the patient was interested in quitting smoking, instead of just outright demanding the patient quit. She inquired about the patients' personal likes and dislikes, which is cool.
- The student was nice and kept the patient informed.
- Very good balance of professionalism and empathy. The appointment moved along quickly but she addressed all concerned and did a lot of listening
- She was thorough with the information collected and provided good advice to the patient.
- The student was formal and polite, but also engaged in social interaction aside from medical evaluation making the patient feel more open and free with providing the information.
- The student was friendly and relaxed. She had an open demeanor, which seemed to make the patient comfortable. She repeated what the patient said in order to make sure she had everything correct.

- The student maintained a constant interaction with the patient. No lack of communication. The student provided the patient with all information relevant to the situation. The student addressed the concerns of the patient in a caring and empathetic way, showing concerns for the patient. The student enforced the patient's decision to come get checked out. The student was very respectful in all dealings with the patient.
- She seemed very easy to talk to. Her demeanor was very friendly, and she seemed to be putting effort into listening to what the patient was saying.
- She was very thorough and caring. She obviously was trying to do a very good job from the very beginning, double checking her questions and making sure she covered every point that she could remember. She tried her best to express understanding and empathy and to come off as casual yet competent.
- The student was good at asking the correct questions, and guiding the patient to help them figure out the answers if the question was confusing (ex. "associated symptoms").
- She was friendly and made the patient feel better by asking about her hobbies and complimenting her.
- The student did a good job of building a relationship with the patient, summarizing the patient's complaint, and explaining the tentative diagnosis.
- I think she did very well. She definitely communicated with the patient in a natural way. She sounded very professional and confident.
- I like how the student did recap the symptoms. I also like how she expressed sympathy about the condition and asked questions about my personal life.
- She did a good job empathizing with the patient, especially towards the end. Additionally, she solicited the patient's thoughts and feelings regarding what she was most concerned about and validated the legitimacy of her concerns, which is good for making the patient feel respected and like she is being listened to. She did a good job summarizing the final diagnosis and the relatively benign nature of the diagnosis, especially the fact that she was not having a MI. Overall, her social and communications skills were pretty good except for the few awkward exchanges. She exhibited and conveyed empathy, respect, and genuine concern for the patient.
- The student related with the patient and was thorough in their questioning to identify what was causing the issue. They were very friendly and encouraged the patient to communicate their wishes.

With respect to interpersonal and communication skills, <u>what could this student DO BETTER</u> in his/her next patient encounter?

Your Feedback:
- Student's Feedback

Package A:
- When the patient revealed that she was a smoker, the student was a bit blunt in asking about quitting.

Package B:
- Little genuine discussion, no summary, and only broad instructions at the end.

Package C:
- Irrelevant line of questions about personal life could be avoided such as talking about relationships.

- I think she did a great job. I am not used to conversational doctors visits so I think she should keep up the great work.
- She said "weird" at one point which doesn't seem like a good thing to say to a patient, you shouldn't say what they are experiencing is weird.
- The student could let the patient go into a little more detail about treatment options.
- She didn't seem to give the patient a choice in treatment, just presented it as 'this is what we'll do'
- Didn't explain medication, but she didn't prescribe it then.
- They could ask if the patient wanted to quit smoking, instead of saying "today" they could have said sometime in the future as this would have been less forceful to the patient
- I think she could have read the patient a bit better. When she suggested cutting back on fried and spicy foods, the patient seemed to not really want to do this.
- The student did not offer other suggestions and was kind of stern in their decision to cut down on spicy food. The student could have offered the patient the choice of cutting down or trying the medication. This is only a minor suggestion, as the student did very well overall.
- There's not much. There were a couple of odd pauses, and a slip of the tongue when she asked about who the patient lived with but overall she was very good.
- She needs to relax a bit more. She was obviously nervous and this made me nervous. She needs to have more confidence, practice her questions more, and take her time more. She seemed to be in a rush at some points.
- When the patient showed hesitation to the dietary changes, the student could have used that as an opportunity to ask the patient what she would rather do. If I had been that patient, I wouldn't have felt like I had the ability to say "no, I don't want to stop eating spicy foods, what's the deal with the meds".
- The medical student shouldn't have said to the patient that heart attack is definitely a possibility. The patient remained scared and felt relief just after the medical student came back and told her that Dr. Harris said it is not a heart attack. It didn't feel like a real reassurance anyways.
- The student could improve by engaging in more collaboration with the patient about which course of treatment the patient wishes to take - it seemed as though the patient was reluctant to make the lifestyle changes suggested as a first-line treatment, but the student didn't give the patient the chance to discuss the pros and cons of the two approaches and weigh in on her preferred course of treatment.
- I don't think she needs to change anything about her communication skills. In fact, if she didn't tell that she is a student I would have mistaken her as a real doctor. Good job!
- I think the advice to quit smoking was offensive and abrupt. The questions about family medical history were late in the conversation which made her seem disingenuous. I also think she would do better if she did not use technical language when referring to the impact of symptoms on other parts of the body.
- Well, there are several aspects which could be improved upon for the next patient interactions. A few of these include: -Trying to make the transition, segue, and flow of the typical/required questions more natural and less awkwardly timed/phrased. For example, in this interaction, the MS2 Additionally, she should work on avoiding prematurely closing before obtaining all the necessary information as premature closure makes it seem like the subsequent questions are an afterthought and/or may lead to a

patient questioning the physician's competency/ability. -Another aspect she should work on is avoiding the tendency to ask leading questions as this frequently biases patients' responses and may result in obtaining incomplete and/or inaccurate information which can consequently hamper accurate and prompt diagnosis and management. -She should work on providing more natural responses that are appropriate and unlikely to make the patient uncomfortable. Specifically, her response to the patient's profession e.g. oh that must be hard was obviously off-putting to the patient as noted by her facial expression and recoiling. Additionally, her phrasing of the illicit substance question needs improvement as it was phrased awkward and somewhat abruptly. Moreover, people do not consider tobacco and ethanol as illicit substances so when the MS2 referred to "other illicit substances" that may have struck the patient in a less than favorable manner. -The order of events/questions requires improvement. Specifically, she should have inquired about associated signs and symptoms earlier and should have inquired about nausea (which she omitted and the patient had to actually mention nausea) rather than other less signs and symptoms that are less likely to be associated with GERD.

- The student communicated that the option to remedy the pain was to stop eating spicy foods and briefly mentioned medication. They should have discussed the available medication and side effects and let the patient decide.

Appendix 21: Sample Physical Exam Skills Feedback Package Presentation

Item	Your Rating	Package A	Package B	Package C
Number of evaluators	1 (You)	1	1	19
1. Clean hands before touching the patient (using hand sanitizer or soap and water)	Student's Rating	Correctly	Obscured	Correctly (63%) Incorrectly (0%) Not Done (0%) Obscured (37%)
2. Visually inspect skin on chest (from collarbone to bra line). Inform patient of visual inspection.	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)
3. Place finger pads over aortic area to palpate for any thrills or pulsations	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)
4. Place finger pads over pulmonic area to palpate for any thrills or pulsations	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)
5. Place ulnar surface of the palm over the tricuspid area to palpate for any thrills or pulsations	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)
6. Place finger pads over the mitral area of the heart to feel the point of maximum impulse	Student's Rating	Correctly	Correctly	Correctly (95%) Incorrectly (5%) Not Done (0%) Obscured (0%)
7. Listen with diaphragm of stethoscope for heart sounds in the aortic area	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)
8. Listen with diaphragm of stethoscope for heart sounds in the pulmonic area	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)

9. Listen with diaphragm of stethoscope for heart sounds in the tricuspid area	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)
10. Listen with diaphragm of stethoscope for heart sounds in the mitral area	Student's Rating	Correctly	Correctly	Correctly (89%) Incorrectly (11%) Not Done (0%) Obscured (0%)
11. Listen with bell of stethoscope for heart sounds in the aortic area	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)
12. Listen with bell of stethoscope for heart sounds in the pulmonic area	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)
13. Listen with bell of stethoscope for heart sounds in the tricuspid area	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)
14. Listen with bell of stethoscope for heart sounds in the mitral area	Student's Rating	Correctly	Correctly	Correctly (89%) Incorrectly (11%) Not Done (0%) Obscured (0%)
15. With the patient lying on his/her left side, listen with bell of stethoscope for heart sounds in the mitral area	Student's Rating	Incorrectly	Incorrectly	Correctly (53%) Incorrectly (21%) Not Done (26%) Obscured (0%)
16. Stands to the patient's right during the entire exam	Student's Rating	Correctly	Correctly	Correctly (100%) Incorrectly (0%) Not Done (0%) Obscured (0%)

Appendix 22: Survey Gizmo Layout for the Student Reactions Study

Page 1: Welcome + Screening

Please use a laptop or desktop computer to participate in this study (do not use a mobile device such as a cell phone or tablet). Your computer will need to be connected to the internet and capable of streaming audio and video content.

Are you a currently a medical student in your M# year at Rush Medical College?
( ) Yes
( ) No

Page 2: Consent

[CONSENT WAS HERE]

Page 3: ICS Ratings

Instructions: Part 1 of 3

You are about to watch two videos (History and Plan; Physical Exam) from a single standardized patient encounter (SPE). In this SPE, actors play the role of medical students ("standardized student").

Please rely on your exposure to the Rush Interpersonal and Communication Skills (ICS) Questionnaire and the Physical Exam Checklist to characterize the performance level of the student in each video.

1. Before watching the video, scroll down and re-familiarize yourself with the questionnaire you will use to provide feedback about the student's skills.

2. Watch the video in its entirety
- *We strongly recommend watching the video in full screen in a new browser tab*: Click the "YouTube" icon in the bottom right corner. After the video begins to play, click the frame icon in the bottom right corner to make it full screen.
- The standardized student's face is blurred in each video. We understand some information about the student's behavior is lost this way, but please do your best with the behavior you *are* able to observe.
- You will see 2-3 camera angles of the same encounter presented simultaneously
- Adjust the sound so you can hear the conversation between the standardized student and patient
- You are not allowed to copy, retain, or distribute this video or video URL

3. After watching the video, fill out the corresponding questionnaire
- You can re-watch any part of the video you want while filling out each questionnaire

- Please finalize your evaluation for each video. You will not be able to navigate back to previous evaluations.

Interview and Plan Video

What is the patient's chief complaint?
( ) Knee pain
( ) Chest pain
( ) Fatigue
( ) Headache
( ) Joint pain
( ) Back pain
( ) Persistent cough
( ) Stomach pain
( ) Trouble sleeping

According to the patient, what seems to aggravate his/her condition?
_____
_____

Briefly summarize the medical student's/doctor's treatment plan.
_____
_____

You may find that the descriptions for each of the performance levels ("1-5") don't exactly describe the behavior of the student; that's OK. You should think of level "5" on the questionnaire as the ideal doctor. Use the descriptions as a guide to help you make a decision about the student's general skill level. For example, if you feel that the student did a little better than the *types* of behaviors described in the level "1" description but a little worse than the *types* of behaviors described in the level "3" description, choose level "2".

[ICSF WAS HERE]

Page 4: PE Ratings

Physical Exam Video

What kind of physical exam was performed in this encounter?
( ) Abdominal Exam
( ) Cardiovascular Exam
( ) Extremities/Joint Exam
( ) Neurological Exam
( ) Respiratory Exam
( ) Vital Signs Exam

Page 5: ICS Feedback Packages

Instructions: Part 2 of 3

Please review the three feedback packages below, which are based on the same videos that you just reviewed:

- The number of evaluators contributing to each feedback package varies
- For ease of review, the feedback packages are presented side-by-side, and each package has a unique color assigned to it
- Your ratings and feedback are provided for your reference only, and you don't have to remain committed to them. For example, you may decide that on second thought, specific ratings/comments from other feedback packages better represent the performance in the video.
- Although you are not expected to review the entire Interpersonal and Communications Questionnaire or the Interview and Plan video again, they are provided at the very bottom of this page for your reference.

- For those feedback packages with more than one evaluator:
    - Items using a categorical scale (e.g., Yes/No) are presented as the percentage of evaluators who selected each option
    - Items using a continuous scale (e.g., 1-5) are presented as: Mean (Standard Deviation)
    - For written comments, bullet points separate comments from individual evaluators

[ICS FEEDBACK PACKAGES PRESENTED HERE]

Type in the number of evaluators/raters that contributed to each feedback package:
Package A: _____
Package B: _____
Package C: _____

[RFQQ-ICS PRESENTED HERE]

The materials below are provided for your reference only. You are not required to review them.

[INTERVIEW AND PLAN VIDEO AVAILABLE HERE]

[ICSF AVAILABLE HERE]

Page 6: PE Feedback Packages

[ICS FEEDBACK PACKAGES PRESENTED HERE]

Type in the number of evaluators/raters that contributed to each feedback package:
Package A: _____
Package B: _____
Package C: _____

[RFQQ-PE PRESENTED HERE]

The materials below are provided for your reference only. You are not required to review them.

[PHYSICAL EXAM VIDEO AVAILABLE HERE]

[PESC AVAILABLE HERE]

Page 7: Student Reactions Questionnaire

Instructions: Part 3 of 3

The three feedback packages you just reviewed came from three different sources. The first two feedback packages, A and B, were from evaluators who commonly review student performance in standardized patient (SP) encounters:

- Package A came from a standardized patient rater in the volume typical of events at Rush Medical College (i.e., ratings from a single SP).
- Package B came from faculty raters (behavioral science faculty for the interpersonal and communication skills component; clinical faculty for the physical exam skills component) in the volume typical of events at Rush Medical College (i.e., a single set of ratings for each component).

Package C, the one with 15+ evaluators, was collected from a crowdsourcing website called Amazon Mechanical Turk (AMT). A crowdsourcing site allows a requester to post small tasks to a large pool of online workers, who complete these tasks in exchange for a monetary reward. In this study, the same video review and rating/feedback exercise you, the faculty, and standardized patient evaluators completed was posted as a series of tasks on AMT. These "crowd evaluators" were provided with instructions on how to review the videos and fill out the **Interpersonal and Communication Skills** and **Physical Exam Skills** Questionnaires (you can view the instructions they were provided by clicking the questionnaire names). The following screening criteria were used to select the crowd evaluators:

1. Must have completed at least 500 tasks on AMT with at least a 97% approval rating
    - This criterion was used to control the quality of crowd evaluators
2. Must be from the US and have visited a US-based physician as a patient or as the guardian of a patient within the last 2 years
    - This criterion was used to ensure that crowd evaluators were 1) patients in the not-too-distant past, and 2) understood the norms and expectations around the behaviors and practices of US-based physicians

There are several advantages to using screened crowd raters to review videos of standardized patient encounters:

- Patient as consumer
    - Some studies have demonstrated that standardized patient encounter ratings by trained raters (e.g., faculty, SPs) do not predict patient satisfaction scores. Perhaps then the most appropriate and relevant judge of physician behavior (especially of interpersonal and communication skills) is the ultimate consumer of that behavior, the patient.
    - Providing future physicians feedback about their behavior is a way for patients to understand and meaningfully contribute to the healthcare delivery system.
- Quality
    - Single ratings are notoriously unreliable, especially for non-cognitive domains such as interpersonal and communication skills. One of the best ways to improve reliability is to add more raters. This is difficult to accomplish due to cost and

time constraints associated with SP and faculty raters, but achievable with crowd raters.
    - Several studies have demonstrated that lay raters or groups of lay raters can accomplish certain tasks just as well as, or sometimes better than, individual experts (e.g., ratings of surgical tasks, estimating calories on a plate of food, rating writing samples)
- Diversity
    - The same physician behavior may evoke different reactions in different patients. Since patients are not robots, their preferences for physician behavior vary. Although this may become evident when students encounter higher patient volumes during clerkships, preclerkship students may not receive enough feedback to understand this natural distribution of patient reactions. Aggregating ratings from multiple crowd raters can help students get a better sense of differences among patient preferences early on.
    - Aggregating ratings from multiple crowd raters can help eliminate biases that may occur in single ratings (e.g., harsh/lenient raters, gender/race preferences)
- Speed and scalability
    - Feedback is most effective when it's delivered quickly. Because of staffing levels, there is often a bottleneck with SP and faculty ratings, therefore decreasing the efficacy of the feedback. Crowdsourcing allows many raters to work simultaneously and at all hours of the day, thereby decreasing the turnaround time (for example, all 15+ ratings for Package C were collected in 2-4 hours).
- Costs
    - The cost per rating is significantly decreased with crowdsourcing. Faculty and SP time is significantly more expensive than crowd rater time. In this study, crowd raters were paid just above the national minimum wage. Additionally, the crowdsourcing marketplace is responsive to compensation changes such that the speed of task completion can be increased by the increasing compensation (e.g., if ratings need to be turned around even more quickly)

After reading the explanation above, please answer the following questions:

	True	**False**
Adding raters typically decreases reliability	( )	( )
Crowd raters generally cost less per rating than SP raters	( )	( )
Crowd raters were recruited from many countries	( )	( )

[SRX WAS HERE]

Page 8: Demographics

[STUDENT DEMOGRAPHICS WAS HERE]

Optional: Please feel free to leave general comments or feedback about this study here.

_____

_____

Page 9: Thank You

Appendix 23: SRX Content Analysis. Advantages Comments by Theme.

Diversity/Volume

- Seeing dissenting opinions is useful. Assuming the crowd-sourced population represents the diversity of the patient population, you are more likely to get good varying feedback.
- not just a single person determining if you pass or fail an SPE
- Better view of how multiple patients would view this potential doctor.
- It would be beneficial to receive more varied feedback. Usually the written feedback I receive is very vague and nonspecific, if I receive it at all.
- Having a breadth of feedback is useful because everyone is so different and has different preferences for how they want to be communicated with. Consequently, I think crowd sourcing would be useful to give a more diverse set of opinions.
- I think the more opinions you have, the better idea you get of how you are doing.
- More opinions,
- This would be so so much better than getting feedback from a single faculty member. I have been in situations where different feedback comes from different sources, which I understand is real because so much of this is an art. But getting a larger number of people's feedback would help validate comments so much more.
- They provide more opinions that are unbiased.
- multiple sources of feedback
- the only reason I would want a crowdsourced system would be to have many people evaluate my interpersonal and communication skills.
- there would be more variety of specific feedback
- The diversity of responses was good too.
- They would provide more information per encounter, so that students may get more feedback to work on and improve on between simulated patient activities.
- additional raters who are real patients
- many opinions
- Being able to note what several people found good or bad about the interaction
- More information to aggregate
- Getting feedback from a larger cohort reduces variability in graders. I think this is the best way to determine communication skills is from a large group rather than receiving only one patients perspective.
- lots of feedback
- It's good to get a feel for how I present as a physician (with the same attitude and behavior) to many patients.
- more opinions
- Get more points of view
- Plurality of opinion
- It gives an unbiased third person's layman perspective that is backed by number of evaluators.
- more feedback
- I think crowdsourcing would be really helpful for evaluating flow of conversation, types of questions ask, language used, connection with the patient, etc. It is helpful to get multiple perspectives on the interpersonal communication aspect of SPEs.

- I think it provides a good diversity to interpersonal feedback.
- more diverse responses
- It is an average of many people's opinions which makes it more reliable than just one person's opinion
- I like that they are from all over the U.S because it lends more diversity in terms of age, geographic location, and culture.
- Multiple perspectives on one issue
- If I receive one negative feedback from an SP I may sometimes ignore it. However, if I receive the same feedback multiple times from a crowdsourced system, I would most likely believe it and use it to improve my interpersonal skills.

Patient as Consumer

- the idea that the consumer (patients) are reviewing the encounter of a future physician. You could receive real-time feedback about how people perceive your ability to communicate effectively with patients regarding their medical concerns.
- Allows non medical people to evaluate your skills.
- Actual fresh patient perspective, not worn down by a script or looking for specific things (i.e. a specific missed opportunity in the script to explore patient questions further),
- These would be people who are probably not involved in the medical industry often and so have the perspective of a real patient, which is probably more accepting of different types of communication styles than simulated patients and faculty, who have a close-minded expectation of how students should interact with patients.
- More representative of patient population we will see as residents and physicians.
- It's good to get a feel for how I present as a physician (with the same attitude and behavior) to many patients. I would be able to adjust my behaviors to appeal to the majority of people in that way.
- Get more points of view from someone more in the role of a patient.
- an additional 'patient-oriented' perspective would be useful
- It gives an unbiased third person's layman perspective that is backed by number of evaluators.
- I also think that it provides a more 'real' evaluation of what a patient would think of an encounter.
- Gives you the perspective of an actual patient who has experienced the healthcare system first hand.
- I like that they are from all over the U.S because it lends more diversity in terms of age, geographic location, and culture.

Bias/Quality

- They provide more opinions that are unbiased.
- Students would not be able to blame a negative evaluation on a 'tough grader' or 'the SP didn't like me'
- more standardized
- outsider perspective
- I had a really terrible faculty behavioral scientist my M1 year who was not at all helpful and who was mildly offensive in her feedback, extremely uncomfortable to be around and

annoying. I would have been so thankful for crowdsourcing then. My M2 year I had a much much better behavioral scientist who was helpful and I trusted to give good feedback. I would have trusted her more than crowdsourcing but crowdsourcing over SP evaluations. Sometimes it seems like we're not even watching the same video.

- Sometimes the evaluations just seem wrong, like I know I did an exam maneuver when they mark that I didn't. Then I can't tell if they just didn't try very hard at the evaluation or they think I did something wrong in which case I want to know how to improve but that feedback is never given.
- Objective
- More honest because less personal
- It gives an unbiased third person's layman perspective that is backed by number of evaluators.
- Impartial and honest feedback since they will never see or interact with the students they are rating
- lessen institutional bias
- accuracy

No Additional Advantages

- No additional advantages – everything above is good.
- I think the biggest advantage was already listed above:
- N/A - you've covered it pretty well.
- none.
- I cannot think of an additional advantage
- None.
- No additional advantages.
- I think all of the major advantages have already been mentioned
- N/A

Cost

- saves time and money
- cost
- and cost less money
- cheaper
- Cost
- cheaper
- cost

Speed

- faster responses
- saves time
- timely
- It would be faster
- more timely feedback
- time-efficiency

Specificity

- Usually the written feedback I receive is very vague and nonspecific, if I receive it at all.
- there would be more variety of specific feedback
- Usually the comments are somewhat helpful but most of the time they are pretty benign and don't give me a whole lot to work with to improve.
- More detailed feedback

Other

- if every school uses crowdsourced system, these reviews may have national standardized potential.
- I think that the crowdsourced system is a good way to review interpersonal skills and to see how students measure up to practicing physicians that people interact with on a professional, healthcare level.
- Physicians see thousands of patients over a career, so being evaluated by a crowd makes sense

Appendix 24: SRX Content Analysis. Disadvantages Comments by Theme.

Accuracy/Quality:

- blurring faces also would leave the door open for inaccurate ratings since eye contact and facial expressions are valuable aspects of patient communication
- The reliability of them assessing what Rush is trying to teach us.
- There may be nuances not noticed in the video, especially with eye contact and the general feel of the room.
- Non medical people do not understand the line of questioning or how questions show be phrased. Sometimes in medicine staff have to be straight forward and patient's may not like the way the question is being asked but it is vital a clear question and answer are formed. For this reason a crowd source will not understand a physician perspective.
- I would question the reliability of the crowdsourcing due to lack of on site training. I would also be concerned about the generalizability of a rating because different persons may evaluate different students.
- Cannot adequately screen crowds for accuracy, crowds do not necessarily know anything about how we are taught to do interviews as med students
- They can't see your facial expressions and they might not understand the situation as well as the actual patient or experienced professional would.
- they are people who aren't in the medical profession
- less reliable/credible
- They definitely do not know the specific details on why some things are done a certain way
- they seemed too positive and i wouldn't be as motivated to improve
- I also thought they graded too easily (lowering the quality of the feedback).
- I think the feedback given by the SP and faculty were much more accurate as opposed to the crowd raters.
- Not trained in how to give feedback
- can't protect privacy without sacrificing eye contact/facial expressionsn by blurring
- may be distracted and only completing for compensation
- random groups of strangers assessing a grade
- I think crowdsourcing would be really hard to keep anonymity and get a good evaluation. I've been told many times that my empathy is apparent in my body language and my facial expressions and that that is a very important part of the encounter.
- Less formal training
- The crowdsourced individuals do not have the same knowledge base as faculty evaluators. Although, I often notice that SPs also only know how to evaluate specific physical exam maneuvers and don't recognize correct alternatives either. The crowdsourced individuals are not actually at the encounter so their experience is not the same as the SPs.
- crowds are not medical professionals
- we are debasing our standards to fit those of an anonymous crowd, rather than holding our medical students accountable to our own high standards
- Judging of actual skills might be poor

- unfamiliar with medical training process and may not be able to give appropriate and fair feedback to students based on level of training
- Lay people are not as familiar with techniques or terms as professionals and might mark something as missing when it was just done differently.
- Student's grades being negatively impacted by raters unfamiliar with the structure and specific requirements for SPE at Rush. For example, when screening for SIGECAPS a lay person may find it weird to be asking seemingly unrelated questions about the patient's personal life whereas an faculty rater would be familiar.
- They are not as well trained in what we are taught to do or what the expectation is.
- depending on how many people evaluate each encounter you run the risk of having a set of mostly poor quality evaluators
- The crowdsourced participants may not take the videos seriously and may not provided constructive feedback.
- For example, we are required to evaluate a patient's social history which includes asking about jobs, alcohol usage, hobbies (exercise), etc. However, in the feedback given in Package C, one comment stated that the student asked unnecessary questions.

Physical Exam Evaluation Apprehension:

- I definitely do not think that they should review physical exam skills as you have no way of knowing their ability to do so.
- Crowd may not be properly educated on certain aspects of physical exam techniques, etc...
- crowds do not know proper physical exam technique
- training - grading on physical exam
- I also cannot trust crowdsourced systems for PE review.
- not trained in physical exam
- not knowledgable about physical exam procedures
- Medical experience would be decreased, so physical exam evaluations could be skewed.
- Probably not great for physical exam because they don't know what is and is not correct.
- I do not know how well the crowdsource would know how to perform a physical exam or the proper technique.
- Also it is difficult to tell if an exam maneuver is performed appropriately from just watching a video of it.
- Not familiar with physical exam manuevers
- I don't think using a crowdsourced system would be appropriate if we were being evaluated on physical exam skills. There are many times when the SP marks a maneuver incorrect, strictly because they were taught one specific way while in clinical practice a maneuver might actually be done multiple ways and not be wrong. I think using crowdsourcing would amplify this problem.
- I do not think that the crowd raters are a reliable judge of if the physical exam procedures are being done accurately since they have never been trained in this capacity.
- I don't think the crowdsourced system would be a good way to evaluate physical exam techniques because it is hard to judge whether a technique is appropriately done just based on video. I think the person who the exam is being done on is truly the best judge of that.

- May not be fully aware of proper physical exam techniques.
- They may not know how physical exams are done and if they are done accurately.

Privacy:

- privacy would be a major concern as crowd sourcers could be potential future patients- i think that students faces would need to be blurred and that they shouldn't state their full name
- Privacy concerns are the biggest disadvantage.
- Would worry about being identified by my institution.
- privacy
- there are also privacy issues
- voice still normal which violates privacy
- privacy concerns for patients and student
- Privacy is a huge issue.
- Privacy, I would not feel comfortable with a video posted of myself
- This isn't really additional, but in terms of privacy, I would even be a little concerned about people recognizing my voice... especially if this became such a common thing that people knew that you could look on this website to see all medical students interviewing SPs.
- Privacy is a big issue.
- I think privacy is the main issue of using a crowdsourced system.
- Privacy issues if the student's faces are not blurred

Other:

- I see no disadvantages, especially if supplemented by faculty and SP ratings. SP rating give additional context on what it felt like to be in the room, and faculty can provide more perspective on protocol.
- Would be much more beneficial to pay senior medical students to complete reviews of student videos in a timely manner than random crowd source.
- Also more anxiety provoking for students.
- Time required for feedback
- collectivism is a major source of evil in the world
- they may not be representative of the patient populations we will see (e.g., how many women would be included in each scenario?)
- There may be weeding of irrelevant/useless evaluations.
- honestly these forms of feedback do very little to motivate me to change how I do things--generally speaking, I do what I want/feels right anyway, so it's not like it'll have much impact one way or the other
- Crowdsourced ratings regarding interpersonal communication skills may seem less important to students compared to feedback from a content expert or a behavioral science faculty member.

Bias:

- Also, crowdsourced systems may put certain students (particularly minority students) at risk for receiving lower scores based on bias from a larger crowd and/or negative (more

importantly, inappropriate) comments from a crowdsourced system. I think very effective screening procedures would have to be in place to ensure that fair, objective individuals were evaluating the students.

- there is bias - certain members of the crowd could hate doctors/have bad prior experiences and thus rate us lower than they should
- If a student has an accent or English isn't their first language, I would bet they would get a lower rating from the crowdsourced system.
- can use personal bias
- Moreover, the people who participate in the crowdsourcing may hold biases

No Disadvantages:

- I see no disadvantages
- N/A - you've covered it pretty well.
- None.

Feedback clarification:

- I think it takes away from students ability to meet with physicians and behavioral scientists to reflect on their performance.
- Difficult to seek clarification if you don't understand feedback or think it is incorrect
- Faculty ratings are important because students can break down an evaluation and discuss minute details face-to-face; this is missing from a crowdsourced system.

Feedback volume:

- Sometimes giving too much information about a patient encounter is not useful. Some of the information given was repetitive or didn't agree with one another. Also if there's too much text feedback, students will be less likely to read through all of it.
- Too many cooks in the kitchen! Would rather have one reliable faculty and SP with a lengthy feedback session in person on how to improve PE skills and interviewing.