2011

When Does Fidelity Matter? An Evaluation of Two Medical Simulation Methods

Nneka Joseph
University of South Florida, nik2334ki@yahoo.com

When Does Fidelity Matter? An Evaluation of Two Medical Simulation Methods

by

Nneka Joseph

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Michael Branncik, Ph.D.
Michael Coovert, Ph.D.
Doug Rohrer, Ph.D.

Date of Approval:
May 3, 2011

Keywords: Performance, Replication, Training,
Assessment, Reliability, Validity

Table of Contents

List of Tables

Abstract

Job or task simulations are used in training when the use of the real task is dangerous or expensive, such as flying aircraft or surgery. This study focused on comparing two types of simulations used in assessments during a Clinical Performance Examination of third-year medical students: computer enhanced mannequins and standardized patients. Each type of simulation has advantages, but little empirical work exists to guide the use of different types of simulation for training and evaluating different aspects of performance. This study analyzed performance scores for different competencies as well as the reliability and validity of the different simulation types. The results showed that though neither simulator was generally superior to the other, the mannequin performed surprisingly well on interpersonal tasks such as communication and history taking.

Chapter One

Introduction

An organization with skilled, knowledgeable workers is usually successful and produces goods and service that are competitive in the market. Training is a key ingredient in developing an effective and efficient staff. It was identified as one of the most pervasive methods for increasing employee productivity and conveying the goals and culture of the organization to new personnel (Arthur, Bennett, Edens & Bell, 2003). Training is also important for current employees' development because tasks, techniques and strategies tend to evolve. Formal training programs are a multi-billion dollar industry in the United States, so employers appear to believe that training is worth considerable expense (Arthur et al, 2003). In medicine it becomes even more imperative that training and assessment are used effectively because training translates not only to a dollar value but also to saving human lives.

In medical training one of the most popular training methods is to use simulations. There have been many empirical studies conducted that assess the merits of simulators comparing them to other training methods. However, very few have explored the differences between different types of simulators. This study is designed to determine whether the fidelity differences between two types of simulations matter. Before introducing the hypotheses and methodology for the present study a review of training is presented which will examine training design and techniques. This would provide

background for the examination of medical training specifically medical simulations. To further under simulations a portion of this paper will introduce the concept of fidelity. Finally the competencies, which are measured by medical training, are presented. Based on the literature review, the importance of research concerning medical simulators is shown and the aim of this study is to provide empirical evidence that can assist medical administrators and faculty.

Effective Training Design

Creating an effective training program is a complicated process which takes into account many factors and stages for development. The first step in developing an effective training program is a needs analysis (Salas and Cannon-Bowers, 2001). For training to be successful, it is essential to identify those individuals that need training (person analysis), analyze the organization structure (organization analysis) and review the job itself (task analysis). The purpose of this step is to identify the task deficits within the organization and determine the knowledge, skills and abilities (KSA) needed for an employee to fill those gaps.

The second stage of developing a training program is to design a training method. Research in this area has discovered that there are four main guidelines that facilitate the design of a successful training method (Tannebaum &Yukl, 1992). The first guideline states that the method should be consistent with the cognitive, physical or psychomotor processes that lead to mastery. In terms of medical training, this refers to consistency within the curriculum. Concepts developed during lecture, information read in text books and the tasks practiced during lab time should be consistent with each other and actual on the job tasks. The next guideline is opportunity for practice. Learners need to actively

practice the behavior being taught as research has shown the "more active the production the greater the retention and transfer of the knowledge acquired" (Tannebaum & Yukl, 1992). Most medical programs have laboratory and clinical hours which allow students to practice and develop their clinical skills. Third, a training method should be adaptable. It should be tailored to match the aptitude and prior knowledge of the trainee. Finally, a training method should incorporate relevant, credible and constructive feedback delivered within a short time period.

The third stage of effective training design is to create a training environment that is consistent with the working environment (Salas and Cannon-Bowers, 2001). Environment includes physical appearance, tone and conditions. For example, if the actual job is in a high stress environment, such as a hospital, then trainees should be exposed to high stress conditions during training. As stated in guideline 2 of designing a successful training method, consistency is an important part of training especially for transfer of learning (Tannebaum & Yukl, 1992).

Training Techniques

Once the steps in training design stage have been followed the next stage is to develop or choose the most appropriate training technique. In the medical field there are many different types of training, including lecture, textbooks, video recordings, structured laboratory experiences such as dissection, patient interview role plays, patient simulators for diagnosis, and on-the-job training. Each type appears to have advantages and disadvantages, some of which are described in this paper. However, there has been little systematic empirical investigation of the merits of the techniques for achieving competency in various domains. Rather, as in most formal instructional settings, the

choice of training technique is left to the best judgment of the instructor. Due to the importance of the outcome of medical training, it would seem worthwhile to provide the instructor with empirical evidence upon which to base his or her judgment.

A similar point may be made with regard to the evaluation of skill or training outcome assessment. Different methods of skill evaluation include paper-and-pencil tests and their computerized descendants (e.g., multiple choice, matching, and fill in the blank), responses to simulated tasks (e.g., reading an x-ray, diagnosing a simulated patient, inserting a central line into a mannequin), and evaluations of care of actual patients as made by expert judgment or objective outcomes. Here, too, the choice of methods is left to instructors or perhaps administrators, who must choose largely on the basis of their own good judgment rather than on the basis of an empirical literature,

Simulations

The broad choice of the instructional method or method of assessment is too large a topic for any single empirical investigation. This study focused more narrowly on simulations used in the assessment of clinical skills in medicine. The aim of the paper was to gather empirical evidence about the advantages and disadvantages of using different types of simulation to assess a common set of skills needed in initial patient encounters.

Medical simulations come in many different forms including simulated devices and standardized patients as well as clinical exercises (Scalese et al, 2007). Simulated devices include individual 3-D body parts (limbs or chest), computer enhanced mannequins (CEM) or more recently, virtual reality computer simulations. These

simulations can mimic multiple medical tasks, and may be designed for many different purposes.

Simulations are useful in medical training and evaluation because trainees can make mistakes in a risk free (to the patient) environment. The increasing use of simulation in medicine is due mainly to changes in the delivery of health care, the reduction of patient availability for education purposes, a need to minimize medical errors and ultimately to focus on patient safety. Research has found that simulations can help with all of these factors as well as demonstrate the skills and competence of medical professionals (Scalese, Obeso, & Issenberg, 2007).

Simulations in medicine are commonly used for assessment during the objective structured clinical examination (OSCE). This examination consists of multiple stations at which students are given specific tasks to perform and are evaluated on the skills needed to handle each task (Newbel, 2004). Typically, many of the stations contain simulated medical cases portrayed by either a standardized patient (actor) or mannequin (CEM). Students spend approximately 5 – 10 minutes at each station. The standardized patient and/or faculty member observer completes an evaluation of the student's performance at the station. The purpose of the OSCE is to demonstrate competence in clinical settings, so that procedural knowledge can be assessed.

The commonly used types of simulations in an OSCE are standardized patients (SP) and computer enhanced mannequin (CEM). A SP is usually a healthy person who is trained to act out a particular medical case or condition. The actor is required to present the case accurately and consistently and then to evaluate the performance of the physician/student that examined them based on specific criteria (Beullens, Rethans,

Goedhuey & Buntinx, 1997). A CEM is programmed to present a medical case and imitate the biological workings associated with the patient's condition. CEMs allow students to interact with the interface such that if they inject a drug, the mannequin would have a reaction such as an increased heart rate or a drop in blood pressure. The important design factor of both types of simulation is that they can be adjusted to meet the needs of the individual. Therefore cases can be presented in varying levels of difficulty depending on the skills of the assessee (Scalese et al, 2007).

Since SPs and CEMs are the most widely used simulation types it is important to explore the effectiveness of these simulators for medical training. This is done in two ways. First, by reviewing studies that has shown the practical aspects for learning (i.e. providing feedback) and secondly, examining the psychometric properties of simulators as an assessment tool.

Effectiveness of SP and CEM

Simulations have been proven to be an effective method for training medical personnel. A literature review of articles dating back 34 years found 10 reasons why simulations are an essential tool (Issenberg, Mcgaghie, Petrusa, Gordon & Scalese, 2005). The three main reasons were feedback, practice and consistency. Fifty-one percent of the articles noted that immediate feedback was important for the success of simulations as a training technique. Feedback is a core part of learning and is needed for retention and transfer of what is taught (Sheull, 1986). Repetition and practice were identified as important for outcomes in 39% of the articles reviewed. Learning occurs when students practice the material because it helps to develop their existing schemas (Rumelhart & Norman, 1981). The review also found that when simulation tasks

incorporate the concepts and information taken from other components of the curriculum, (e.g. lecture and textbooks) students improve their overall test scores. Consistency with all aspects of training was previously mentioned as an important component of training design (Tannebaum & Yukl, 1992).

Additionally, simulations were shown to develop teamwork skills and to provide students with a total understanding of the patient care system (Issenberg et al., 2005). Many medical errors are due to problems in the system due to lack of teamwork (Bogner, 1994). The work shift rotations cause medical teams to be constantly changing; therefore it is imperative to train personnel to adjust to new teams quickly and efficiently. Simulations can be used to facilitate teamwork training and thus improve the safety of patients.

The effectiveness of SPs to evaluate medical skills was examined in a study that involved the collaboration of five universities. There were 36 simulated medical cases portrayed by SPs that were used for assessment. The results of the evaluation identified specific students with skill deficiencies, who were not detected using other methods (paper/pencil examination). These findings also revealed the strengths and weaknesses of each school's curriculum (Stillman et al, 1990). This study demonstrated the effectiveness of SPs to differentiate among students and to show deficiencies in student abilities as well as school curriculum which are not exposed by other assessment methods.

A similar study was conducted to assess the clinical skills of 134 primary care physicians. Clinical skills were defined as history taking and preventative screening items. The assessment involved 17 different cases that were presented by standardized

patients and scored using checklists. The results show that on average 59% (SD= 8%) of the history items were asked by the participants. Physicians missed questions about the patients' past history such as past medical illnesses, asking women about their last menstrual cycle and other symptoms that may not be directly related in the patients' view, such as night sweats. In terms of preventative screening, physicians consistently asked about smoking and alcohol use but not about non-injection recreational drugs. They also tend to shy away from inquiring about patients' sexual habits such as condom use, number of sexual partners and history of STDs (Ramsey, Curtis, Paauw, Carline & Wenrich, 1998). The study demonstrated that using SPs can identify the strengths and weaknesses of participants in terms of their clinical skills. This type of assessment was able to pinpoint exactly which types of questions were omitted and what types of errors in diagnosis could arise as a result.

In another example, nursing students' clinical skills and competence were measured using the Objective Structured Clinical Examination (OSCE) to assess the effectiveness of scenario-based simulation training using CEMs. A pre-test/post-test design was used such that students were tested at the start of the study and then 6 months later. The experiment consisted of an experimental group exposed to simulation training for two afternoons and a control group that had no exposure to simulators. The results showed that while both groups improved their test scores the experimental group had significantly higher scores (Alinier, Hunt, Gordon & Harwood, 2006). The implication of these findings is that using simulation training with lecture can effectively improve students' proficiency and skills.

Assessment of Simulations

The next aspect of determining the usefulness of simulations is to examine the construction and validation of them as measurement instruments. There have been many studies designed to evaluate the psychometric properties of the OSCE. One study examined the internal consistency reliability of the OSCE by analyzing data from several years of the examinations (1995 – 1999) using 26 clinical cases. The results included internal consistency estimates for history taking skills ($\alpha = 0.68$), physical examinations ($\alpha = 0.53$) and communication skills ($\alpha = 0.76$). The dimension reliability scores were a bit low except for communication, apparently because calculations were based on a small number of items. However, the overall examination reliability ranged from $\alpha = 0.72 – 0.88$ (Brailovsky & Grand'maison, 2000).

The inter-rater reliability of OSCE was examined by another study that used 18 stations and two examiners per station. The data were gathered from examinations completed over the time period 1997 – 2000 in New Zealand. The average inter-rater correlation over the four examinations was $r = .78$ (Wilkinson, Frampton, Thompson-Fawcett & Egan, 2003).

Criterion validity was examined in a study that analyzed the psychometric properties of an OSCE for senior pharmacy students. Validity was determined by calculating the correlations between scores from the OSCE and those measured using a written exam and clinical marks. The global ratings included skills of empathy, coherence (organization and focus), verbal skills, nonverbal skills, and overall impression (knowledge and skills integration). The findings showed that when professional SPs

played the role of patients, concurrent validity was $r = .44$ with written exam and $r = .23$ when compared to scores from clinical mark (Sibbald & Regehr, 2003).

Authors have also investigated the reliability and validity of individual simulations that comprise an OSCE. A review article investigated the pros and cons of different types of medical assessment including clinical simulations, specifically standardized patients. It was mentioned that one of the pros of clinical simulations is that internal consistency reliability may be as high as $\alpha = 0.85$ to 0.90 (Epstein, 2007). Another review of several articles found that researchers consistently reported inter-SP-reliability of about .85 (Beullens et al, 1997).

In terms of validity, medical assessment using SPs have been described in the medical literature in four different ways. When patient cases are developed they are based on curriculum content and evaluated by subject matter experts (faculty) to ensure that they include the skills required of medical students. Such cases are described as content valid. Response process validity is defined as "evidence of data integrity such that sources of error associated with the test administration are controlled or eliminated to the maximum extent possible" (Downing, 2003). SPs were described as response process valid because their evaluations are usually based on checklists which have been shown to reduce or control evaluation errors (presumably because the items in the checklist are readily observable behaviors such as washing hands). Item response models and generalizability theory both provide evidence for what is described as internal structure validity. SP items are usually free of differential item functioning (DIF), a type of bias, and the samples of behavior in SP cases can be generalized to the behaviors displayed on the job (Downing, 2003). SP assessments have also been described as valid because they

have statistical relationships with other variables (what we would call criterion related validity). For example, SPs scores for history taking and physical examinations skills have been correlated with clinical competence ratings (r = .60) (Swartz, Colliver, Bardes, Charon, Fried & Moroff, 1997).

CEMs have also produced evidence of reliability and validity as assessment tools for medical student evaluation. They have been shown to produce ratings that have good internal consistency and inter-rater reliability. For example, Boulet et al. (2003) used CEMs to imitate cardiac and respiratory responses for anaphylaxis (an allergic reaction that causes shortness of breath), asthma and pneumothorax (air is present in the chest causing fainting and shortness of breath). The participants were scored by medical faculty members on their overall performance based on physically examining the patient. The results showed good internal consistency reliability score, $\hat{\rho}^2 = .74$ (Boulet et al, 2003). $\hat{\rho}$ This study also found the inter-rater reliability was $\hat{\rho}_{xx} = .97$ (both estimates from Boulet et al are generalizaiblity coefficients; see Shavelson & Webb, 2006).

The validity of CEMs was investigated by examining the relationship between simulator scores and number weeks spent in critical care electives. The hypothesis was that relevant, content related experience would be positively associated with performance outcomes. The results showed a significant positive relationship ($r = .24$, $p < .05$). The CEM assessment was also capable of differentiating between types of participants (resident or student) in terms of their experience and knowledge level (Boulet et al, 2003).

Fidelity

In order to study simulations it is imperative to understand the concept of fidelity. This is because medical simulations can vary in appearance from individual 3-D body parts (limbs or chest) to sophisticated virtual reality simulators. They can also differ in terms of their function or purpose (e.g. to practice surgical techniques or to practice giving bad news). As stated before, medical simulations are typically used to duplicate the physical form of the human body, to imitate the biological workings, and to create circumstances that medical personnel will encounter on the job. The quality of imitation, that is, how well the simulation duplicates the human body, is commonly referred to as fidelity. Fidelity is the degree of similarity between the simulation and real situation it is designed to imitate (Hays & Singer 1989).

At one time there was a view among researchers that increasing the fidelity of a simulator would result in increased transfer of learning. As stated before the effectiveness of a training method is measured by the trainee's ability to transfer his/her newly acquired knowledge and apply it to the real world (Tannebaum & Yukl, 1992; Salas & Cannon-Bowers, 2001). This view was based on the findings that simulations demonstrated better transfer when compared to books and lecture and so it was assumed that the more realistic the training method, the more transfer of learning would occur (Alessi, 1988).

However other researchers thought that this explanation was too simple for a construct as complicated as learning. In terms of learning, there are many factors to consider when designing an effective training method, such as the level of the student being taught (e.g. novice versus expert). Some studies (e.g. Miller, 1974; Dittrich, 1977)

found that simply increasing the fidelity of the simulation can lead to diminishing returns. For novices, low fidelity would be best because too many cues and information can be distracting and confusing, resulting in little or no learning. As students' knowledge increases, then the fidelity should also increase but only to a certain point. For example there is evidence that an expert pilot can learn and transfer his/her learning to a real plane if trained on a medium to high fidelity simulator (Alessi, 1988).

A meta-analysis examined the specific characteristics of a simulator that leads to transfer of learning (Hays, Jacobs, Prince & Salas, 1992). The analysis focused on flight simulators and effects of fidelity. It was discovered that that simulation training is task specific and is only effective for certain tasks. Fidelity of the simulator should be high for tasks of interest and low for others. For example, motion cues are not necessary for all tasks involved in flight training.

Fidelity is more complex than determining if it should be high or low. There are two dimensions of fidelity which are physical fidelity and functional fidelity (Hays & Singer 1989). Physical fidelity refers to the actual equipment and if it looks like the real thing. Functional fidelity is the extent to which the simulator acts like real equipment (Allen, Hays & Buffardi, 1986). To demonstrate how fidelity dimensions can affect trainee performance the following study is summarized.

Allen, Hays, & Buffardi (1986) conducted a study to investigate the effects of fidelity dimensions on transfer of learning and also to determine individual difference variables, (i.e. logical capacity, analytic ability and general interest) that may interact with fidelity. The task was to troubleshoot an electronic device by determining which electrical relay or pull-up panel was faulty. In this study physical fidelity was

manipulated by variations in the representation of different components and their spatial relationships. High physical fidelity was the actual equipment and the medium level was a simulator designed to match the size and appearance of the equipment but some of the components were represented with either pictures or wooden knobs that did not move. The low physical fidelity simulator did not look like the real reference system, the pull-up panels, relays and output devices were represented by labeled rectangles connected by lines which corresponded to wires in the system. The degree to which informational feedback was provided to the participants via the equipment was functional fidelity. In the high level participants were given status information about the components and the output device. The medium level only provided information about the components where as in the low level no information was given.

The results of the study found that for physical fidelity only time taken to solve the problem was significant, such that higher fidelity resulted in less time. This possibly occurred because participants had to take a little more time to orient themselves to the equipment in the lower fidelity groups. For the other dependent variables such as number of tests used to solve the problem or number of repeated tests performed there was no main effect for physical fidelity. Functional fidelity was found to have a main effect for all of the dependent variables measured. Higher fidelity groups demonstrated better transfer of learning. There was only one significant interaction and that was for the number of repeated tests. Those in the high physical/high functional and high physical/medium functional conditions repeated fewer tests than other conditions. The study also included a control group that completed the tasks without any training. When compared to the trained groups, across all training conditions, trainees were able to

complete the tasks quickly, with fewer tests and more accurately than the control group (Allen et al, 1986).

The conclusion that can be drawn from this study is that the dimension of the fidelity and the degree of its manipulation can have an impact on the training outcome. Physical fidelity effects performance only when time is an important factor but functional fidelity affects most aspects of task performance. What this means for evaluating medical simulations is that all aspects of simulator fidelity would have a significant impact on case design and the performance of trainees. Therefore administrators would need more assistance when choosing a simulator beyond using intuition.

Dimensions Assessed

With the intention of providing a comprehensive review of medical training it is necessary to describe the skills that are being taught and assessed during the OSCE. In encounters between doctors and patients, the two main competencies typically required of health professionals are clinical competence and interpersonal and communication skills. Clinical competence refers to the skills involved in patient history taking and physical examination. Interpersonal skill is the ability to" interact with a patient during a clinical encounter" (Colliver, Swartz, Robbs &Cohen, 1999).

Clinical competence is an important skill that is necessary for the job as a medical professional. Taking a thorough history report about medication, allergies and prior illnesses provides valuable information that can prevent some medical complications. Trainees should practice interpreting an ECG or x-ray, checking the body for wounds or injuries and listening to internal organs such as the heart and lungs. A doctor should also demonstrate their technical skills such as inserting a needle, stitching an open wound or

performing surgery. Clinical competence directly translates to saving lives and reducing medical errors.

There are several benefits associated with doctors being effective communicators. The first is it makes patients feel more comfortable while answering questions. This could lead to identifying the patient's problem more quickly and accurately. Secondly, patients report a satisfactory experience with the communicative doctor and they leave with a better understanding of their problems and treatment options. Thirdly, patients are more likely to heed the doctor's advice on behavioral changes as well as follow their treatment regimen. There is evidence that patients' are less stressed, anxious or depressed when they develop a rapport with their attending physician (Maguire & Pitceathly, 2002).

Medical personnel should be trained in the skills needed to be an effective communicator. For example, establishing eye contact at the beginning and throughout the consultation indicates to the patient that the doctor is listening and interested. During consultation a doctor should encourage the patient to give precise details, such as dates when symptoms started and their feelings. Another skill is "active listening" which incorporates responding to key words, restating information provided by the patient and being receptive to any corrections or misunderstandings. Doctors should inquire about how the illness is impacting the patient's life, such as work, social activities and family life. Addressing the patients' concern early in the consultation and recommending reading material will help patients learn more about their problem and alternative methods of treatment. As part of their consultation, doctors should include the patient in the treatment processes by discussing treatment options and determining the best course of action for him/her. It is important to mention lifestyle changes that would lead to

effective treatment. Finally, empathy demonstrates to the patient that his/her feelings are taken into account and provide encouragement to cope with the situation (Maguire & Pitceathly, 2002).

Non-verbal communication is also a significant part of communication skills. A physician's facial expression, tone of voice and body language can influence responses from the patient (Mast, 2007). It is also essential for physicians to be able to read the non-verbal cues of their patients. In many instances patients may be embarrassed to talk about their illness or problems, or they may lie about how much pain they are in. Physicians need to be able to observe body language and facial expressions which may provide more information necessary for diagnosis. Also patients may express their emotional states through non-verbal clues which provide information about psychological and social issues (Levinson, Gorawara-Bhat & Lamb, 2000).

The Present Study

In the literature many comparisons have been made between simulated patients and other methods of training such as multiple choice, oral and written examinations, global rating scales, medical records reviews and patient management problems. However, little has been done comparing two different simulation methods. The literature review conducted has also shown that the fidelity of training matters, in terms of type and degree which seems to have a profound effect on learning and the transfer of knowledge to the job. It was also noted that training using simulations is task specific and that it is important to determine that the desired behaviors are being taught when designing and using one. This study compared two simulations, CEM and SP, used in assessing clinical competence and communication. The research examined performance

mean score differences, the reliability and validity of measures taken in comparable scenarios using both simulations for the same group of examinees in order to provide an empirical evaluation of the methods. The results of this study provide data about when fidelity matters, and useful evidence for instructors to aid in deciding what type of simulation to use.

CEMs and SPs were compared to determine which one is better suited for presenting the case so that students can demonstrate the skills being assessed. The study also compared these simulations by analyzing the reliability and validity of each simulation for each dimension. The data for the current study was obtained in a Comprehensive Clinical Performance Examination (CPX) which is a specific type of an OSCE. The CPX consisted of 12 stations but for this study the focus was on the patient with a pneumothorax played by a SP and represented using a mannequin. The data was collected as part of Frederick. R. B. Stilson's dissertation entitled 'Psychometrics of OSCE Standardized Patient Measurements' (2008).

The type of fidelity was specifically functional fidelity. These two simulations were considered to have high physical fidelity because SPs were actually people therefore they would physically match and the CEM used in the CPX were designed to represent a human body in appearance. The literature also showed that functional fidelity had a greater effect on learning and transfer (Allen, et. al., 1986).

Performance Mean Scores

Communication skills incorporate both verbal and nonverbal cues and though the CEM has a person talking for it, the medical students do not receive any visual signals such as facial expression. It is a challenge to establish eye contact, show empathy and

read facial expressions which are all behaviors associated with communication skills. Communicating with a talking mannequin may seem very artificial and therefore reduce the amount of communication between the student and patient. Because of this, it is proposed that the SP has more functional fidelity in terms of this dimension.

Hypothesis 1: Communication skills are better elicited when a standardized patient is used compared to a CEM which will be shown by higher communication scores for the SP than the CEM.

For the competency physical examination the CEM is considered to have higher functional fidelity because it is actually designed to imitate the biological workings of a human. For this case the patient has a pneumothorax and the CEM is programmed so that a student hears a collapsed lung. A SP is a healthy person acting the part of a sick person, therefore when a student listens to the SP's lungs, the student does not hear the collapsed lung, but rather a relatively healthy lung. This can be a little confusing and coupled with the SP's own health problems, the student may make some errors during this process.

Hypothesis 2: CEM physical examination scores will be better than those earned with the SP.

Because taking a patient's history is a necessary part of an effective evaluation and it is simply a series of questions that a student must ask a patient, for this dimension each type of simulation has the same level of functional fidelity. Therefore there should not be any differences between the two. The student is hypothesized to ask the appropriate questions whether encountering a CEM or a SP.

Hypothesis 3: For history taking, the effect size (Cohen's d) for the difference between the two methods will be less than 0.2.

Reliability

In previous research the CEM and SP have been shown to provide reliable and valid measures of clinical competence and communication skills but they have not been compared to each other to determine if one is more reliable or valid. This study examined these measurement variables for each types of simulation to determine if one is more reliable or valid over the other. In terms of reliability there are two forms that are analyzed, internal consistency and inter-judge reliability. The items on the checklist are examined for internal consistency by calculating a . Cronbach's alpha for each dimension of the scale.

Hypothesis 4: The alpha level for each dimension would be good ($\alpha \geq .8$) for both simulation types.

The second type of reliability is inter-judge reliability. There were scores from five raters for each dimension for each simulation. The inter-judge reliability is computed for each dimension such that; for all dimension by comparing the ratings of the five raters.

Hypothesis 5: Inter-rater reliability scores would also be in a good range but physical examination ratings will be more consistent for CEM and SP reliability level would be higher for communication. Inter-judge reliability is not significantly different for history taking skills.

Validity

Criterion validity was also examined in this study. The case was developed using medical professionals as SMEs, designed using curriculum material and the standards used for the OSCE, which establishes content validity. The criterion validity however has been looked at in only a few studies. Cases presented in the OSCE measure clinical judgment, diagnostic reasoning, treatment decisions and communication skills (Epstein, 2007). The criterion was performance in terms of scores on the dimensions communication, clinical competence and knowledge summed over the remaining exercises in the OSCE.

Hypothesis 6: The pneumothorax case represented by both the CEM and SP would have criterion validity when it is compared to the scores from the other CPX stations. It is hypothesized that the CEM has stronger correlations for the clinical competence dimensions and SP has a stronger correlation for communication skills. They both have positive correlations with history taking skills.

The main point of having students perform these tasks is to predict their performance as a medical professional. Therefore it is important to determine if performance on one CPX station can predict performance on the other stations.

Hypothesis 6(a) is that a regression would show that John Long scores significantly predict scores obtained from the other CPX stations for physical examination, that John Sexton communication scores would predict the communication scores from the other stations and that they would both predict history taking scores.

Chapter Two

Method

Participants

The participants for this study consisted of third year medical students from a University in the southeastern United States. These students participated in an assessment involving a patient with a pneumothorax presented both by a standardized patient and a CEM. At this medical school, students are required to participate in this assessment and so data was used from students who went through the process over the past two years. The identities of the SP and the students were kept anonymous and confidential since they were not included in the analysis.

Material/Procedure

For this study the focus was on the case involving a patient with a pneumothorax (collapsed lung). The case was either presented by an SP, John Sexton, who was the victim of a stab wound or by a CEM, John Long, who acquired his injury from a biking accident. Prior to their examination of the patient, the students were given a medical chart that provided information about the patient as would ordinarily be available through intake. The students were required to take the patient's history, conduct a physical examination and discuss a recovery plan with the patient. Students were assessed on their history taking skills, conducting an appropriate physical examination and correct diagnosis of the injury as well as their communication skills.

The standardized patients were trained and paid for their work in the CPX. The training for all SPs consisted of regular training sessions which took place throughout the year in which they learned their job requirements via lecture and slides and then played the role for the physicians before they were certified to be an SP. SPs were accepted based upon the demographic needs of the case and typically play only one case. During the actual examination, SPs had access to their respective scripts until a student entered the room. This allowed them to be as consistent as possible. For this study, the case was played by three different SPs.

The checklist for this case was broken into 3 main sections, interpersonal/communication skills, history taking and physical examination. The first section, communication, included active listening skills (e.g. the student listened actively and heard my concerns), showing empathy (e.g. the student demonstrated understanding and compassion), being considerate (e.g. the student considered my feelings as well as my concerns) among 10 other items. These were rated on a 1-5 scale such that 1 meant poor, student has major weakness in this area and 5 represented excellent, the student could not be better. The ratings on these questions were equivalent to points on the test therefore if a student got a rating of 4 then they earned 4 points towards their final grade on this assessment.

The checklist continued with a series of 12 yes/no questions about the student's history taking behavior (e.g., did the student elicit or allow you to volunteer information concerning when the injury occurred?). For every question that the patient selected "yes" the student earned 1 point. The final section evaluated the student's physical examination technique. It was also a sequence of yes/no questions each earning 1 point for every

"yes" response (e.g., did the student wear gloves during the examination?). The checklist ended with two questions that did not count towards the student's grades but were valuable in terms of the having a better understanding of the students' performance. The first was; "what is your overall impressions of this student's performance?" This question was rated on a 1 – 5 scale where 1 was not able to assess this student's performance and 5 meant outstanding impression. The second question was open-ended which asked the patient to provide any additional comments on the student's performance. There were slight differences in the questions asked between the JL and JS cases and only the questions that matched each case were used.

The students also had a series of questions to complete after examining the patient. This consisted of 6 open-ended questions regarding the case (e.g. Please write your exact diagnosis of this patient. List 5 potential concerns/injuries based on this patient's presentation). A faculty member or trained administrator graded the student's responses based on a rubric designed for the case. The scores on each dimension for a station were weighted and combined with scores on the other stations to arrive at an overall final grade.

Each video recording of the role-play was viewed by five additional raters, at least two raters for each simulation type. One of the raters was a graduate student in industrial/organizational psychology; the others were either graduate or undergraduate students. All of the additional raters obtained the same training as the SPs and used the same CPX evaluation forms. Based on observations from a live CPX and discussions with the SPs, it was determined that SP ratings do not discriminate in the category of communication. Students were consistently rated as a 4 or 5 and little distinctions were

made. Also SPs did not pay specific attention to being asked about their family, or the students showing empathy although these items were on the checklist. Therefore, the raters were additionally trained on rating the communication portion of the evaluation form using a Behaviorally Anchored Rating Scale (BARS). The BARS was created by interviewing several SPs in order to determine what behavior needed to be exhibited by the medical student in order to receive a certain rating. For example, in order to score 3 out of 5 on the introduction, the medical student had to make eye contact with the SP. In order to receive a 4, they additionally had to shake the SP's hand. Behavioral referents were created for each communication item. Each rater saw between 258 and 282 videos.

Analyses

The aim of this study was to compare two simulations; CEM and SP, in terms of performance mean scores, the reliability and validity of scores provided for assessing clinical competence and communication. Each student was rated at least twice (once by two raters) and their average score for each item was computed. This was done for each simulation type and for each dimension.

Performance Mean Scores

The students were exposed to both patients, John Long the CEM patient and John Sexton the case played by a standardized patient and so a paired sample t-test was used to analyze within person differences. The dependent t was used to test hypothesis 1 that students have higher communication scores with the SP than with the CEM. This statistical method was used to analyze hypothesis 2 which stated that the CEM group had higher physical examination scores. It was also used to test hypothesis 3 that for history taking the effect size is less than 0.2. In addition to the t-test, for each comparison the

effect size, Cohen's *d* was calculated to determine the strength of the relationship between type of simulation and performance score.

Reliability

To test for differences in internal consistency, the average response to each item across judges was computed for each dimension for each type of simulation. Coefficient alpha was computed by dimension by type of simulation (i.e., one alpha for communication, history taking, and physical examination separately for the CEM and the SP). Differences in alpha were tested by dimensions across simulation types (e.g. a significance test was computed to compare communication internal consistency between the CEM and the SP).

Intraclass correlation (ICC) was used to compute the inter-rater reliability between judges for each dimension. The rater scores were compared for each dimension for each simulation type. A total of 6 (5 raters) correlations were computed. To test the differences between the ICC for the simulations the Feldt test was used. This produced 3 *W*-test scores which determined if the CEM had a higher inter-rater reliability for physical examination and to establish if SP assessment had higher reliability for communication skill. It was hypothesized that history taking skills would not be significantly different between these simulation types.

Validity

Finally to compute criterion validity, the correlation between the pneumothorax case scores with clinical competence from the other cases was calculated. A Pearson r was used to determine if there was a relationship between scores from the pneumothorax case with those from the other CPX cases (e.g. Rachel Brown - abdominal pain

(appendicitis)) for each dimension. A high correlation between the scores on this case (John Long/John Sexton) with the scores from the other case would indicate that this measurement method is a valid measure of communication skills and clinical competence. Hypothesis 6 was that CEM and SP would have criterion validity but CEM would be more strongly correlated for the physical examination dimension and SP would have higher validity with communication. They would be both positively, significantly related to history taking scores. Linear regression was used to test hypotheses 6(a) to determine if performance on the pneumothorax case predicts performance on other medical cases.

Chapter Three

Results

The station scores from131 students were used to calculate mean comparisons between John Long (mannequin) and John Sexton (standardized patient). Only common items for each competency were used which resulted in 6 items for communication, 5 for history taking and 2 for the physical examination (see Appendix A).

Performance Mean Scores

Contrary to hypothesis 1, the results of the *t*-test showed that for the communication competency there was no significant difference between presentation types. John Sexton's mean score (M = 20.29, SD = 2.60) was not significantly higher than John Long's mean score (M = 20.01, SD = 2.44; t (130) = -1.05, p = 0.29; d = 0.11. The results for physical examination were in the opposite direction predicted in hypothesis 2, such that John Sexton' scores had a significantly higher mean score (M = 1.12, SD = 0.63) than John Long's mean score (M = 0.65, SD = 0.82), t (130) = -6.02, $p < 0.05$. This comparison showed a medium effect size d = 0.64. The third hypothesis was also not supported as the John Long history taking scores (M = 3.95, SD = 0.83) were significantly higher than John Sexton's (M = 2.63, SD = 1.12); t (130) = 12.60, $p < 0.05$ with an effect size (Cohen's d = 1.34) that exceed the hypothesized value of d= 0.2.

Table 1. Correlations between scores for John Long and John Sexton

	John Sexton		
John Long	Communication	History Taking	Physical Exam
Communication	**0.26**		
History Taking	-0.02	**0.27**	
Physical Exam	0.00	0.13	**0.23**

N = 131 **Bold** indicates $p < 0.05$

Reliability

Hypothesis 4 that for each presentation type alpha levels for all competencies would be at or above $\alpha = 0.8$ was partially supported. John Long communication internal consistency was $\alpha = 0.8$ and physical exam was $\alpha = 0.96$ but the history taking competency was very low $\alpha = 0.37$. For John Sexton history taking reliability was also low $\alpha = 0.40$, communication had a score of $\alpha = 0.83$ and physical exam Cronbach's alpha was 0.77.

Inter-judge reliability was calculated using two-way mixed intraclass correlation (ICC) based on the scores from 21 students who were rated by 5 raters. For communication, the John Long ICC was 0.78 and the John Sexton ICC was 0.83. The history taking correlations were 0.92 for Long and 0.94 for Sexton. For physical examination, John Long's ICC was 0.94 and John Sexton's ICC was 0.73. These results are consistent with the hypothesis that there would be good inter-rater reliability with

values above or around 0.8. However for hypothesis 5 which postulated that raters will be more consistent for John Sexton for the communication competency while raters would have more agreement for John Long for physical examination was partially supported. The Feldt test (Alsawalmeh & Feldt, 1994) for comparing alpha coefficients was used and it showed that only physical examination had significant differences such that raters were more consistent when scoring these items with John long than with John Sexton; $W = 4.5$, $p < 0.05$ (See Table 2 for all reliability results). All other alpha values were not significantly different from each other.

Table 2. Reliability Results for John Long and John Sexton by Competency

		Competency		
Type	Reliability	Communication	History Taking	Physical Exam
John Long	Cronbach α	0.80	0.37	0.96
	ICC	0.78	0.92	0.94
John Sexton	Cronbach α	0.83	0.40	0.77
	ICC	0.83	0.94	0.73

Cronbach α: N = 131 ICC: N = 21, Raters = 5

Validity

For validity analyses scores from two other CPX cases were used. Rachel Brown and Samantha Browning are two female SP roles that portrayed the same case, which involved abdominal pain (appendicitis). Correlations and regressions were calculated to determine if John Sexton and John Long competencies related to the corresponding skills

being measured in the female cases. It was also important to determine if competencies in the male cases could predict competency scores for the female cases. Student scores for all four cases were used. Calculations were based on the set of common items among the cases. For history taking items only 4 questions were used in all four cases.

For communication, scores on the John Long scenario were significantly related to scores on the Samantha Browning scenario ($r(129) = 0.44$, $p < 0.05$) but not with score on the Rachel Brown scenario. John Sexton also had a significant correlation with Samantha Browning ($r(129) = 0.37$, $p < 0.05$) and also with Rachel Brown ($r(129) = 0.25$, $p < 0.05$). The only significant correlation for history taking was between Long and Brown ($r(129) = 0.21$, $p < 0.05$). John Sexton was significantly correlated with Samantha ($r(129) = 0.19$, $p < 0.05$) and Rachel ($r(129) = 0.23$, $p < 0.05$) for physical examination. There were no relationships for John Long (see Table 3 for complete correlation results)

Table 3. Correlation Results for John Long and John Sexton with other Cases

Competency		1 John Long	2 John Sexton	3 Rachel Brown	4 Samantha Browning
Communication	1	1			
	2	**0.26**	1		
	3	0.12	**0.25**	1	
	4	**0.44**	**0.37**	0.17	1
History taking	1	1			
	2	**0.27**	1		
	3	**0.21**	0.09	1	
	4	0.14	0.16	**0.33**	1
Physical Examination	1	1			
	2	**0.23**	1		
	3	0.04	**0.23**	1	
	4	0.11	**0.19**	**0.22**	1

N = 131 **Bold** indicates $p < 0.05$

Six regression models were analyzed, 2 for each competency. Each female case was used as the dependent variables (DV) and both male scenarios, John Long and John Sexton, were entered in the model together as independent variables (IV). For communication both models were significant. When Rachel Brown was the DV, John Sexton was the only significant predictor. For Samantha Browning both IVs were significant predictors of the case's communication scores. The first history taking model

was significant such that John Long predicted scores in the Rachel Brown case. The model with Samantha Browning as the DV was not significant. For the physical examination competency when Rachel Brown was the DV the model was significant and John Sexton successfully predicted scores for this case. The second model was not significant despite the significant beta weight for John Sexton.

Table 4. Regression Results for significant models for John Long and John Sexton with other Cases

Competency	Case Models		B	SE b	β	t	P
Communication	RB	JL	0.05	0.08	0.06	0.71	0.48
		JS	0.19	0.07	0.24	2.68	0.01
	SB	JL	0.33	0.07	0.37	4.74	0.00
		JS	0.23	0.07	0.27	3.49	0.00
History Taking	RB	JL	0.21	0.09	0.20	2.30	0.02
		JS	0.05	0.07	0.07	0.78	0.44
Physical Exam	RB	JL	-0.01	0.43	-0.01	-0.11	0.91
		JS	0.29	0.11	0.23	2.66	0.01

N = 131 Key: JL = John Long; JS = John Sexton; RB = Rachel Brown; SB = Samantha Browning

Chapter Four

Discussion

Medical simulations are an important part of training for medical professionals. Simulations come in many forms but the most popular have been CEMs and SPs, which have been used interchangeably. The purpose of this study was to determine whether the fidelity differences between two types of simulations would matter. Based on the literature review it was determined that the focus of this paper would be on medical simulations specifically standardized patients and computer enhanced mannequins. Although previous research has evaluated the effectiveness and psychometric properties of this training method in comparison to other options (i.e. lecture) this study provided empirical evidence that would facilitate administrators when using deciding between different types of simulators. In the subsequent sections, the results are further summarized and interpreted, followed by a discussion of practical implications, study limitations, and future directions.

Summary and Interpretation of Results

The results of the study showed that neither SP nor CEM is uniformly superior to the other and that both showed evidence of reliability and validity. However, the CEM performed better than expected at eliciting interpersonal behavior from the students.

Performance Mean Scores

The analyses showed that there was no difference between presentation types for the competency of communication. It was hypothesized that because John Long was a mannequin, communication would seem more artificial than with a live person and therefore the scores would be lower. As previously described the voice for John Long is an actual person in another room and therefore communicating with the mannequin was simply like talking to someone in another room. Students also performed tasks such as shaking the mannequin's hand and looking it in the eyes, hence, communicating with the patient was not as artificial as originally proposed and thus the hypothesis was not supported.

The students scored significantly higher with John Sexton than with John Long for physical examination and this could be as a result of the questions which were asked for this competency. To correctly perform this task students had to listen to the patient's lung (or heart) both at the front and back as well as under the gown. The mannequin is to some extent difficult to move in order for the student to get to the patient's back. Students could simply ask the SP to lean forward and lift his shirt in order to perform the task appropriately.

For history taking it was found that students had higher scores when they interacted with the mannequin than with the standardized patient. It was proposed that because this task is a series of standard questions to be asked of all patients there would be no difference between the presentation types. Upon further examination of the items used for this competency it was found that students generally did not ask the SP about their past medical history and associated symptoms. It is possible that because the actors

who played these roles were young men, it did not occur to the students that there was a need to ask about their medical history. The mannequin is not age specific and therefore there was no bias when it came to asking these questions.

Reliability

Overall the reliability was within convention such that they were either above or at 0.8. For internal consistency, physical examination and communication Cronbach's alpha values ranged from 0.77 – 0.96 for both presentation types. However history taking had values lower than 0.4 and this could be as a result of number of items and the nature of the questions. These questions asked about different aspects of someone's medical history which by statistical definition does not indicate homogeneous items though theoretically they relate to the competency of interest. For inter-rater reliability the ICCs were all in a good range, 0.73 – 0.94. Generally speaking, neither the SP nor CEM resulted in more reliable measures. However, it was found that for physical examination, raters were more consistent for John Long than John Sexton. This is possible because it was clear to the rater if the student performed this task themselves with John Long but with the SP there may be some discrepancy in scores if the actor assisted the student by lifting their shirt or leaning forward without being asked.

Validity

It was hypothesized that the scores from the male cases would predict similar competency scores for the female cases. The results showed that when Rachel Brown was the dependent variable in the model that for each competency the model was significant. However there was no distinct male case that was consistently a significant predictor. For communication and physical examination John Sexton predicted Rachel

Brown scores while John Long was a predictor for history taking scores. When Samantha Browning was the DV only the communication model was significant and both male cases were significant predictors of those scores. There is no clear explanation for these results. It may be on account of John Long having better history taking scores and John Sexton having significantly higher physical examination scores. Students did equally well on the communication competency with each case and this may explain the male cases both predicting Samantha Browning's scores.

Implications

The results of this study provide implications for the testing of medical students using the OSCE method. The outcome of the simulation comparisons suggests that the different presentation types do elicit different performance scores from the students. This is useful for administrators when determining which simulation is used for skill assessment. One suggestion would be to consider the constraints of the simulator when creating items for evaluation. This was seen with the physical examination competency which required students to move the CEM. Due to the weight of the mannequin students were unable to perform the task of listening to the lungs/heart at the back but it was not an accurate reflection of their aptitude on this skill. Therefore more diligent evaluation of the simulator itself is needed before creating test items.

Comparisons on this competency also brought to light another issue which may have to be addressed through training of SPs and using external raters. From inter-rater reliability findings as well as score difference for physical examination there is the possible explanation of SPs assisting students by lifting their gown or leaning forward. Based on video observations and conversations with SPs and administrators it appears

that SPs develop bonds with the medical students, sincerely wishing them well, and are apparently providing some assistance during the evaluation process. Since this is an important component of this skill it would have to made clear to SPs that they are not to provide any assistance in this area. It may also be prudent to have external raters to identify instances where the student may have been assisted by the actor. SPs interact with many of these students on different occasions throughout the CPX, which may facilitate the bonding that is being seen. Therefore administration may want to limit the amount of contact that SPs have with students.

The most interesting finding of this study was that for interpersonal skills, the mannequin performed equally well and even better than the standardized patient. Students were able to interact with the mannequin for their communication tasks, even shaking the CEM's hand, which was unexpected. For history taking students were even more inclined to ask questions of the mannequin than the SP. This demonstrates that it is not intuitive which simulator would be better for any particular competency and empirical evidence should be the determining factor when deciding which simulator should be used for medical assessment. This study has shown that further research in evaluating different simulation types is needed in order to more accurately evaluate the skills of medical personnel.

Limitations

As with any study there are a few limitations. One limitation for this study would be the number of items that were used to analyze the differences between presentation types. Each case originally consisted of about 27/28 questions. For Communication there were 12 different items for John Long and 7 for John Sexton. History taking items

were 11 for CEM and 12 for SP and for the physical examination competency students were to perform 4 different tasks for John Long and 9 for John Sexton. The analyses were performed on only the common items and this resulted in only 6 communication items, 5 history taking and 2 physical examination items. This reduced number of items could have affected the comparisons that were performed.

Another limitation was that students participated in the CPX at two different time points during the academic year. Students were tested using John Sexton during the Fall and John Long in the Spring semester. During the time between tests students were exposed to other training such as lectures, gaining more medical knowledge (reviewing text book material) and being exposed to other types of simulations and this could have an impact on the findings of this study.

Future Research

This study highlighted two main types of fidelity, physical and functional fidelity but there is a third dimension known as psychological fidelity. Psychological fidelity, concerns the extent to which the trainee identifies the simulation as a believable substitute for the trained task (Beaubien & Baker, 2004). This type of fidelity does not specifically refer to the equipment itself as with the other two dimensions but it takes into account the perception of the trainee. Future research should include this fidelity dimension in its assessment since this can have a significant impact on students' performance between the two presentation types. Beaubien and Baker (2004) also proposed that it is the interaction among the three fidelity dimensions that makes the difference in training performance and this interaction should also be explored.

It would also be prudent to broaden the criterion validity analysis. Medical students participate in many simulation training exercises beyond those mentioned in this study. They are tasked with varying assignments such as giving bad news to terminal patients or working with patients with more complicated medical conditions. Future studies could determine if performance in the CPX could predict performance in these other training exercises and if it could predict overall performance of third year medical students when all tests and examinations are included.

Conclusion

Overall this study has found several differences in measures based on similar cases portrayed by a standardized patient and by a computer enhanced mannequin. Most of the differences were not in the predicted direction, however. Some differences appear due to the physical attributes of the mannequin, and some have no obvious explanation. An important result was that the measurement of student competence in communication appeared not to be degraded using the mannequin.

References Cited

Alessi, S.M. (1988). Fidelity in the design of instructional simulations. *Journal of Computer Based Instruction, 15,* 40 – 47.

Alinier G., Hunt B., Gordon R. & Harwood C. (2006). Effectiveness of intermediate-fidelity simulation training technology in undergraduate nursing education. *Journal of Advanced Nursing 54*, 359–369

Allen, J.A., Hays, R.T. & Buffardi, L.C. (1986). Maintenance training simulator fidelity and individual differences in transfer of training. *The Journal of Human Factors and Ergonomic Society, 28,* 497 – 509.

Alsawalmeh, Y.M. & Feldt, L.S. (1994). A modification of Feldt's test of the equality of two dependent alpha coefficients. *Psychometrika, 59,* 49 - 57.

Arthur, W. Jr., Bennett, W. Jr., Edens, P.S. & Bell S.T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88,* 234–245.

Beaubien, J.M. & Baker, D.P. (2004). The use of simulation for training teamwork skills in health care: how low can you go? *Quality & Safety in Health Care, 13,* i51 – i56.

Beullens, J., Rethans, J.J., Goedhuys, J. & Buntinx, T. (1997). The use of standardized patients in research in general practice. *Family Practice, 14,* 58 - 62.

Boulet, J.R., Murray, D., Kras, J., Woodhouse, J. McAllister, J. & Ziv, A. (2003). Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology, 99*, 1270–80

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

Colliver, J.A., Swartz, M.H., Robbs, R.S. & Cohen, D.S. (1999). Relationship between clinical competence and interpersonal and communication skills in standardized-patient assessment. *Academic Medicine, 74*, 271 – 274.

Downing, S.M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education, 37*, 830–837.

Epstein, R.M. (2007). Assessment in medical education. *New England Journal of Medicine, 356*, 387-96.

Hays, R.T., Jacob, J.T., Prince, C. & Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis. *Military Psychology, 4*, 63 – 74.

Issenberg, S. B., Mcgaghie, W.C., Petrusa, E.R., Gordon, D.L. & Scalese, R.J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher, 27,* 10–28.

Jacobs, J.W. & Dempsey, J.V. (1993). Simulation and gaming: Fidelity, feedback and motivation. In J.V. Dempsey & G.C. Salas (Ed.), Interactive instruction and feedback (pp.197 – 227) Englewood Cliffs, NJ: Educational Technology Publications Inc.

Levinson, W., Gorawara-Bhat, R. & Lamb, J. (2000). A study of patient clues and physician responses in primary care and surgical settings. *The Journal of the American Medical Association, 248,* 1021 – 1027.

Luck, J., Peabody, J.W., Dresselhaus, T.R., Lee, M. & Glassman, P. (2000). How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. *The American Journal of Medicine, 108,* 642 - 649.

Maguire, P. & Pitceathly, C. (2002). Key communication skills and how to acquire them. *British Medical Journal, 325,* 697-700.

Mast, M.S. (2007). On the importance of nonverbal communication in the physician–patient interaction. *Patient Education and Counseling, 67,* 315–318.

Murray, D.J., Boulet, J.R., Kras, J.F., McAllister, J.D. & Cox, T.E. (2005). A simulation-based acute skills performance assessment for anesthesia training. *Anesthesia & Analgesia, 101,* 1127–34

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

Ramsey, P.G., Curtis J.R., Paauw D.S., Carline, J.D. & Wenrich, M.D. (1998). History-taking and Preventive Medicine Skills among Primary Care Physicians: An assessment Using Standardized Patients *The American Journal Of Medicine,104* 152 - 158

Salas, E. & Cannon-Bowers, J.A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52,* 471 – 99.

Scalese, R.J., Obeso, V.T., & Issenberg, S.B. (2007). Simulation technology for skills training and competency assessment in medical education. *Journal of General Internal Medicine, 23,* 46–9.

Shavelson, R.J. & Webb, N.M. (2006). Generalizability theory. In J.L. Green, G. Camilli & P.B. Elmore (Ed), Handbook of complementary methods in education research (pp. 309 – 322). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

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

Stillman, P.L., et.al. (1990). An assessment of the clinical skills of fourth-year students at four New England medical schools. *Academic Medicine, 65,* 320 – 326.

Swartz, M.H., Colliver, J.A., Bardes, C.L., Charon, R.C., Fried, E.D. & Moroff, S. (1997). Validating the standardized-patient assessment administered to medical students in the New York City consortium. *Academic Medicine, 72,* 619 – 626.

Tannenbaum, S.I. & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology, 43,* 399-441.

Vu, N.V. & Barrows, H. (1994). Use of standardized patients in clinical assessments: recent developments and measurement findings. *Educational Researcher, 23*, 23-30.

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

Appendix

Appendix A. Common Case Items

Communication Items
1. **Introduction**: introduced self appropriately
2. **Confidence**: Displayed confidence
3. **Comfort**: Made me feel comfortable
4. **Listened**: Listened actively and heard my concerns
5. **Consideration**: of MY gestures, body language, and feelings
6. **Conclusion**: concluded with a brief overview and treatment plan

History Taking
1. Asked how long you have had your current symptoms?
2. Asked about associated symptoms?
3. Asked about your past medical history?
4. Asked what medications you are on?
5. Asked about tobacco or alcohol use?

Note: RB & SB did not have item 3.

Physical Exam				
John Long		John Sexton	Rachel Brown & Samantha Browning	
1. Listened to your lungs correctly (3 places, under gown)	i.	Listened to your lungs on both sides with deep breaths	i.	Listened to your lungs on both sides with deep breaths
	ii.	Listened to your lungs under gown	ii.	Listened to your lungs under gown
			iii.	Listened to your lungs right middle lobe
2. Listened to heart correctly (in 4 places with diaphragm, under gown)	i.	Listened to your heart	i.	Listened to your heart
	ii.	Listened to your heart under the gown	ii.	Listened to your heart under the gown