IMPACT ON THE PSYCHOMETRIC PROPERTIES OF AN OBJECTIVE STRUCTURED CLINICAL EXAM FOR THIRD YEAR PHARMACY STUDENTS USING FIRST YEAR STUDENTS AS STANDARDIZED PATIENTS

by

Debra J. Sibbald

A thesis submitted in conformity with the requirements for the degree of Master of Arts
Department of Adult Education, Community Development and Counseling Psychology
Ontario Institute for Studies in Education of the
University of Toronto

© Copyright by Debra J. Sibbald 2001
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.
IMPACT ON THE PSYCHOMETRIC PROPERTIES OF AN OBJECTIVE STRUCTURED CLINICAL EXAM FOR THIRD YEAR PHARMACY STUDENTS USING FIRST YEAR STUDENTS AS STANDARDIZED PATIENTS

Master of Arts
2001
Debra Joy Sibbald

Department of Adult Education, Teaching and Learning
Ontario Institute for Studies in Education of the University of Toronto

Abstract

Significant changes in the clinical context of health professional education and impressive research on the measurement of clinical competence support the Objective Structured Clinical Exam (OSCE) as the preferred means of performance-based assessment and standardized patients as an adjunct to current teaching and evaluation methods. While there is research to support the use of standardized actors (SPs) as patients, using students as patients and raters has been little studied across disciplines and not in pharmacy. This study investigated the impact of using first year pharmacy students as standardized patients. It examined psychometric properties of reliability and validity using junior student, SP and faculty raters to evaluate performance in a senior candidate OSCE. The major findings indicated that using junior students is reliable and valid and may provide financial and learning benefits to participants. Further inquiry into quantitative and qualitative questions appears warranted to capture valuable, generalizeable insights.
Acknowledgements

I gratefully acknowledge the assistance and support of the following people in contributing to this research project:

- My thesis supervisor, Peter Gamlin, for his wisdom, guidance and thoughtful suggestions

- My thesis reader, Glenn Regehr, for his encouragement, support, direction and tutelage into the study of psychometric analysis of performance-based assessment methods and tools

- Dr. Cleo Boyd, who developed the Global Rating Scales used as the measurement instrument in the study, for her valuable insights into formative and summative performance-based assessment

- Dr. Heather Boon, professor of the first year pharmacy social administration course, and model of collegial collaboration, for believing in the potential learning benefits in this work, mandating the participation of juniors in this study as a requirement for her course. prompt and organized assistance in the sorting of juniors for appropriate participation in this study, and encouragement and feedback throughout

- Dr. Brain Hodges, for sharing his clinical research experiences and insights

- Diana Tabak, for her humour, contribution and guidance in the training and use of standardized patients

- For faculty members who contributed to this project

- My husband, Gary Sibbald, for his respect and understanding, and for his insights into health profession education and life-long learning
• My children, Matthew, Timothy and Cathryn who were unfailing in their support, motivation and feedback

Finally, I would especially like to recognize the 0T3 class of junior students in the Faculty of Pharmacy, University of Toronto, for their participation in this study, who so willingly gave their time, energy, enthusiasm and insights to this project. I hope that they too learned from this research experience.
# Table of Contents

Abstract .......................................................................................................................... ii
Acknowledgements ........................................................................................................ iii
Table of Contents ............................................................................................................ v
List of Tables .................................................................................................................. vii
List of Figures ................................................................................................................ i
List of Appendices ......................................................................................................... x

Chapter 1: Introduction ....................................................................................................... 1
  1.1 Background to the Study ........................................................................................... 1
  1.2 Problem Statement .................................................................................................. 3
  1.3 Rationale for the Study ........................................................................................... 4
  1.4 The Purpose of the Study ....................................................................................... 5
  1.5 The Research Questions ......................................................................................... 5
  1.6 Terms and Definitions ............................................................................................ 7
  1.7 Overview of Thesis Chapters ................................................................................ 14

Chapter 2: Literature Review ............................................................................................... 15
  2.1 Introduction .............................................................................................................. 15
  2.2 Performance-based Assessment ............................................................................. 15
  2.3 The Objective Structured Clinical Exam (OSCE) ................................................ 18
  2.4 Assessment Tools: Checklists and Global Rating Scales ...................................... 19
  2.5 Use of Standardized Patients ............................................................................... 23
  2.6 Value of SPs for Teaching and Evaluation of Clinical Skills ................................. 25
2.6.1 Value of SPs to Students ............................................................. 26
2.6.2 Value of SPs to Teaching ............................................................ 29
2.6.3 Value of SPs to Assessment ....................................................... 31
2.6.4 Value of SPs to Educational Programs ...................................... 33

2.7 Technical Issues of the Logistics of Using SPs .............................. 35

2.8 Fidelity of Assessment with SPs: Issues of Realism, Reliability and Validity ............ 36

2.9 Test Security ................................................................................. 41

2.10 Comparison of SPs vs Faculty as Raters ................................... 42

2.11 Use of Students as Standardized Patients .................................. 44

2.12 Impact of Participation on SPs .................................................. 46

2.13 Summary .................................................................................... 48

Chapter 3: Research Design and Methodology .................................... 50

3.1 Description of the OSCE Used in the Study .................................. 50

3.2 Quality Control of Consistency in Role Delivery .......................... 51

3.3 Description of the Data Collection ................................................. 51

3.4 Method for Recording Incidents .................................................. 55

3.5 Case Design ................................................................................ 55

3.6 Description of Assessment Tool: a Global Rating Scale ............... 57

3.7 Description of the OSCE Practices ................................................ 57

3.8 Description of the Study Participants ......................................... 59

3.8.1 Description of the Candidates .................................................. 59

3.8.2 Description of the Standardized Patients (SP) .......................... 59

3.8.3 Description of the Junior Students (JR) .................................... 60

3.8.4 Description of the Pharmacist Examiners (TAs) ....................... 61

3.9 Confidentiality Issues .................................................................. 61

3.10 Ethical Review .......................................................................... 62
List of Tables

Table 1.............................................. Background Evolution of Oral Examination

Table 2a.................................................. The OSCE Circuits Day One

Table 2b.................................................. The OSCE Circuits Day Two

Table 3.................................................. Patients and Raters for the OSCE

Table 4............. Correlations of TA scores between Stations with SPs and JRs as Patients

Table 5.............................. TA Scores with SPs as Patients Compared to Other Measures

Table 6.............................. TA Scores with JRs as Patients Compared to Other Measures

Table 7.......... Correlations of Scores between Stations when SPs and JRs were Raters

Table 8.............................. SP Scores Compared to Other Measures

Table 9.............................. JR Scores Compared to Other Measures

Table 10.............................. TA/SP and TA/JR Scores for Stations 1&2
List of Figures

Figure 1............................................................................. Bloom's Taxonomy of Educational Objectives

Figure 2.................................................................................. Higher Order Critical Thinking

Figure 3.................................................................................. Pictorial Representation of the SOLO Taxonomy

Figure 4.................................................................................. Pyramid of Professional Behaviours
List of Appendices

Appendix One.......................... Summary of Instrumentation (Data Analysis) to be Used

Appendix Two..............................................................Template for Development of Test Cases

Appendix Three.................................................................Global Rating Scale
Chapter 1: Introduction

“In what may be called the natural method of teaching the student begins with the patient, continues with the patient, and ends his studies with the patient, using books and lectures as tools, as means to an end.” (Sir William Osler, 1904, p329)

1.1 Background to the Study

At the Faculty of Pharmacy, University of Toronto, the author, a pharmacist, coordinates two sequential courses in non-prescription medications for large classes of undergraduate pharmacy students. One of the goals is to prepare the student to assume the role of a pharmacist who will accept accountability for community patient care by identifying, preventing and resolving problems relating to self-medication. This involves daily communication with patients, requiring accurate and concise evaluation of their global needs. A high level of skill in oral dialogue, interpersonal interactions and assessment is required to competently implement this professional responsibility. During these dialogues, the pharmacist is expected to efficiently recall, understand, and apply cumulative knowledge in order to efficiently analyse, synthesize and evaluate relevant information. Whether that goal has been achieved should be reflected in the type of evaluation at the course conclusion. Methods of final assessment in other courses at the Faculty consisted solely of written, non-cumulative examinations. In order to accurately address the achievement of course objectives and credibly evaluate each student’s proficiency, a final, cumulative oral clinical skills examination following the completion of the two courses, was developed and implemented, using standardized patients.
The format of the examination evolved with changes over the first four years. The first year tested the feasibility in a pilot project wherein each student had a single ten-minute interview with a simulated patient portrayed by a pharmacist. Both the role-playing pharmacist and an observing pharmacist rated the candidate. One practice was held prior to the exam between the candidate and a pharmacist teaching assistant (TA). Students requested a more realistic portrayal of the patient and an opportunity for more than one test interview.

In year two, each candidate had two interviews, and the patient role was played by a professional actor. The actor and the observing pharmacist rated the candidate. The practice format was changed to a small group interaction where the candidate experienced self, peer and instructor evaluation. The focus of the sessions was contemplation, discussion and rehearsal. Emphasis concentrated on behavioural skills, using the reflective student-centred approach espoused by Boud. (Boud, 1985) Student feedback indicated the value and need for additional practices. In year three, in addition to these group practices, thirty volunteer junior first year students were recruited to role-play a patient in an interview with a senior student, immediately following each class in second and third year. Juniors were selected in the hopes that their lack of therapeutic training would provide insights more closely parallel to the experience of the average consumer, and to increase their awareness and skills as future candidates. The overwhelmingly positive feedback from all students involved was instrumental in allowing this initiative to be mandatory the following year. In year four, the junior class of 120 students was divided into four cohorts. Two were assigned to the role-playing class practices with the second and third year students. A third cohort conducted individual private practices with a candidate (SR) prior to the exam, where the focus was feedback from the juniors about organization and communication, rather than content. The
fourth cohort underwent standardized training to achieve consistency in both role-portrayal and use of the assessment tool. They were then added to the oral examination patient pool, to enable each candidate to have four interviews with a standardized patient: two with a junior student and two with a professional actor. This background evolution of the structure of the oral examination and the formative practices led to the formal study which is the basis of this thesis. (See Table 1)

Table 1: Background Evolution of Oral Examination (OSCE):

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Tests</th>
<th>Patient</th>
<th>Raters</th>
<th>Individual Practices</th>
<th>Class Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>TA</td>
<td>2 TAs</td>
<td>SR with TA</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>SP</td>
<td>SP + TA</td>
<td>1 group (TA with 6SR)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>SP</td>
<td>SP + TA</td>
<td>1 group (TA with 6SR)</td>
<td>16 in 3rd yr (JR with SR)</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2 SPs</td>
<td>SP + TA</td>
<td>1 group (TA with 6SR) + 1 private (JR with SR)</td>
<td>16 in 3rd yr (JR with SR) + 15 in 2nd yr (JR with SR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 JR</td>
<td>JR + TA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1.2 Problem Statement

The Objective Structured Clinical Exam (OSCE) is a performance-based assessment that is gaining wide-spread recognition as an examination format for assessing clinical
competence of health professionals. While it is accepted as a valid and reliable method for the assessment of clinical skills in medicine and nursing, little psychometric information is available on the use of OSCEs in pharmacy. While there is research to support the use of standardized actors (SPs) as patients, the impact of using students as patients and raters has been little studied and only using senior students in junior candidate courses or exams. (Barnes et al, 1977; Harris et al, 1990; Davis Feickert et al, 1992; Papadakis et al, 1997; Ursatine et al, 1996 and Sasson et al, 1999) To date, no investigator has used pharmacy students as standardized patients.

1.3 Rationale for the Study

No study has examined the use of junior students as standardized patients for senior student examinations. As subjects, junior students who have not studied diagnosis or therapeutics, closely represent the unknowledgeable patient. As raters, such students, having taken a communication course, may have a heightened awareness of desirable communication skills. As learners, they may perceive a benefit in terms of application to performance in senior years. Only one study has compared junior student raters to faculty or SP raters, with good inter-rater reliability using checklists for procedural skills, but in this study the students functioned as only observers, not patients. (Bullock G et al 1999)

Positive correlations with other measures of student performance would make this innovation a model for teaching that can also easily be feasible not only in other schools of pharmacy but in the education of other health professionals. As a method that reduces costs; generates a database of valid, consistent, reviewed and successful cases; and provides an experiential venue that is more practical than outside placements in bridging the gap between
learning and readiness to practice; it would be transferable to other providers of health professional education.

1.4 The Purpose of the Study

The purpose of this study is to use the third year self-medication OSCE at the University of Toronto, Faculty of Pharmacy to study the impact of using first year junior students as standardized patients. This will be done by examining the psychometric properties of reliability and validity of using first year student, SP and faculty raters to evaluate performance.

1.5 The Research Questions

Reliability:
The study will examine the reliability of the exam scoring when juniors are used as standardized patients in terms of effects on TA raters, patient raters and candidates’ performance.

- Is there an effect on the scores generated by TA raters when junior students are used as patients as compared to professional standardized patients (SPs)?
- Is there a difference in the scores generated by patient raters when junior students are patients compared to SPs?
- Is the performance of candidates affected when junior students are used as patients compared to SPs?
Validity:

The study will also examine if the oral clinical skills exam is a valid measure as compared with other performances.

These research questions are more specifically outlined as follows: (See also Appendix One)

1. Is there an impact on **Faculty Raters (TA) Scores of Candidate Performance** when using junior students vs. SPs?
   
   a. **Reliability**
   
   i. Is there a difference in TA scores using junior students vs SPs when correlations between stations are compared?
   
   ii. Is there a systematic influence on candidate scores when the means of TA scores are compared for juniors vs SPs

   b. **Concurrent validity:**
   
   i. Do TA average scores using juniors as patients show correlations with other measures* of performance (*other oral exam, GPA's from present and current years, traditional science courses, exit experiential course mark the following year)?
   
   ii. Do TA average scores using SPs show correlations with other measures* of performance (*other oral exam, GPA's from present and current years, traditional science courses, exit experiential course mark the following year)?

2. Is there an impact on **Patient Rater Scores of Candidate Performance** when using junior students vs. SPs

   a. **Reliability**

   i. Is there a difference in patient scores - juniors vs SPs when correlations between stations are compared?

   ii. Is there a systematic influence on candidate scores when the means of juniors scores vs SPs scores are compared?

   b. **Concurrent validity:**
i. Do the junior scores show correlations with other measures of performance? (*other oral exam, GPA's from present and current years, traditional science courses, exit experiential course mark the following year)?

ii. Do the SP scores show correlations with other measures of performance? (*other oral exam, GPA's from present and current years, traditional science courses, exit experiential course mark the following year)?

3. An additional analysis for concurrent validity can be made which compares scores generated by TAs and patient raters between stations 1 and 2:

   a. Is there a correlation in the SP vs TA scores for stations 1 and 2 and the junior vs TA scores for stations 1 and 2?

1.6 Terms and Definitions

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andragogy</td>
<td>Management of learning</td>
</tr>
<tr>
<td>Attitude</td>
<td>A settled opinion or way of thinking, or a behaviour reflect this (Oxford Dictionary, 1998)</td>
</tr>
<tr>
<td>Attribute</td>
<td>A characteristic quality ascribed to a person or thing (Oxford Dictionary, 1998)</td>
</tr>
<tr>
<td>Behaviour</td>
<td>The observable pattern of actions of a person; moral conduct (Oxford Dictionary, 1998)</td>
</tr>
<tr>
<td>Checklist</td>
<td>Content checklists are commonly written by qualified practitioners or examiners who</td>
</tr>
<tr>
<td>Clinical Problem Solving</td>
<td>Ability to process and evaluate information and appropriate action relating to a patient-specific case</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Competency</td>
<td>A cluster of knowledge, skills, attitudes and attributes that affect performance of the job, and that can be measured against accepted standards.</td>
</tr>
<tr>
<td>Communication</td>
<td>The art and technique of using words effectively in imparting one's ideas. To convey information about, make known, to reveal clearly. To have an interchange, as of ideas. To express oneself in such a way that one is readily and clearly understood (Nelson, 1997)</td>
</tr>
<tr>
<td>Concurrent Validity</td>
<td>Establishment of the meaning of test scores in terms of other measures of performance (Colliver &amp; Williams, 1993)</td>
</tr>
<tr>
<td>Construct Validity</td>
<td>Establishment of the meaning of test scores in terms of relationships between candidate</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Performance on the test and other characteristics implied by the construct (e.g., clinical competence) (Colliver &amp; Williams, 1993)</td>
<td></td>
</tr>
<tr>
<td>Content Validity</td>
<td>Establishment of the meaning of test scores in terms of content measured (Colliver &amp; Williams, 1993)</td>
</tr>
<tr>
<td>Continuing Education</td>
<td>A systematic effort to provide education beyond formal education and initial entry level into a profession (Belanger, 1997)</td>
</tr>
<tr>
<td>Empathy</td>
<td>The power of fully comprehending by identifying oneself mentally with a person or object of contemplation (Oxford Dictionary, 1998)</td>
</tr>
<tr>
<td>Empowerment</td>
<td>Provided with the means, opportunities, authorization or license necessary for independence or self-assertion (Oxford Dictionary, 1998)</td>
</tr>
<tr>
<td>Exam Candidate</td>
<td>A student who undergoes assessment of competency for program or course evaluation</td>
</tr>
<tr>
<td>Generalizability</td>
<td>Reproducibility of results across other formats</td>
</tr>
</tbody>
</table>
Stability of test scores over time and across content domains (Regehr, 1999)

Global Rating Scale

An assessment tool to evaluate process skills, typically designed on a 5 point scale which may rate the candidate in separate domains such as empathy, organization and focus, verbal and non-verbal skills, or overall impression of the interview.

Junior Student (JR)

A first year pharmacy student taking basic science or social administration courses who has no therapeutic training

Interpersonal Skills

Facility in relating to relationships between people

Leadership

An aptitude which enables one to be followed by others

Learning

Knowledge acquired by study, act or process

Mentor

Description of a non-parental, competent and trustworthy figure who consciously accepts personal responsibility for the significant development and growth of another individual

Mentoring

A behavioural activity that refers to the
one-to-one relationships that evolves through reasonable distinct phases between the mentor and the adult learning

**Needs Assessment**

Refers to any systematic approach to collecting and analyzing information about the educational needs of individuals or organizations. Educational needs can further be described in terms of knowledge, skills, attitudes and attributes or as levels of competencies. (Moore and Cordes 1992)

**Networking**

Occurs through using contacts to extend information and to achieve goals

**Nonverbal Communication**

Exchange of information without speaking, using gestures, physical expressions or body posture, or written words

**OSCE**

Objective Structured Clinical Examination

**OTC Advisor**

A pharmacist qualified to globally assess a patient who wished to self-medicate in terms of the patient, disease and drug-related concerns, the need for treatment and to give expert advice on treatment options, use and effects

**Performance-based Assessment (PBA)**

Evaluations methods that require the
candidate to demonstrate specific skills and competencies, to apply the skills and knowledge they have mastered (Stiggins 1987)

**Pharmacist**
A professional qualified and licensed to prepare and dispense drugs and to give expert advice on their use and effects

**Predictive Validity**
Establishment of the meaning of test scores in terms of future performance (Colliver & Williams, 1993)

**Professionalism**
Refers to the essential outcomes or ideals that identify and define a profession (Houle 1980)

**Professional Practice**
The knowledge, skills and attitudes (competency) required to deliver the outcomes that define a profession

**Problem-Based Learning**
Inquiry started from given conditions: acquisition of knowledge and skills directed through solving experiential, contextual or situated circumstances

**Reflective Practice**
Competency in delivering professional skills acquired through characteristic deep thought, assessment or evaluation
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>The reproducibility or consistency of testing (Colliver &amp; Williams, 1993)</td>
</tr>
<tr>
<td>Senior Student (SR)</td>
<td>A third year pharmacy student who is a candidate for the oral examination</td>
</tr>
<tr>
<td>Skill</td>
<td>The demonstrated ability to apply knowledge and understanding to perform a task (Ontario Society for Training and Development)</td>
</tr>
<tr>
<td>Expertness; practiced ability; facility in an action; a specific aptitude of a particular type of dexterity (Oxford Dictionary, 1998)</td>
<td></td>
</tr>
<tr>
<td>Standardized Patient (SP)</td>
<td>A normal individual who is taught to simulate every aspect of a patient’s illness in a totally consistent manner. Often, professional actors are used (Barrows &amp; Abrahamson, 1964)</td>
</tr>
<tr>
<td>Teaching Assistant (TA)</td>
<td>A pharmacist who is a trained teaching assistant</td>
</tr>
<tr>
<td>Teamwork</td>
<td>The combined action of a group when effective or efficient (Oxford Dictionary, 1998)</td>
</tr>
<tr>
<td>Validity</td>
<td>Accuracy of a prediction or inference made from a test score (Cronbach, 1971)</td>
</tr>
</tbody>
</table>
Establishment of the meaning of test scores in terms of content measured and relationships between examinee performances on the test and other characteristics implied by the construct (e.g., clinical competence) (Colliver & Williams, 1993)

Verbal Communication

Exchange of information by speaking

1.7 Overview of Thesis Chapters

Chapter 2 reviews and synthesizes the literature that supports and informs the study of the use of standardized patients and students in OSCEs used to evaluate professional competency, and the fidelity of the assessment tools. Chapter 3 describes the research methodology and design used for gathering and analyzing the data to explore the research topic. Chapter 5 presents the summarized research findings. Chapter 6 provides an analysis and interpretation of the findings, a summary of the conclusions and implications for both practice and further theoretical exploration.
Chapter 2: Literature Review

2.1 Introduction

To develop a framework for the thesis, this chapter presents a selected literature review. It begins with an overview of performance-based assessment, followed by an outline of the objective structured clinical examination as the preferred form of PBA. It will contrast the two typical assessment tools used in assessment, the checklist and the global rating scale. The literature on the use of standardized patients will be reviewed in terms of value to students, teaching, assessment and program development. Technical issues of logistics will be articulated, fidelity issues of realism, reliability and validity will be summarized, test security will be examined, and comparisons of faculty vs SP raters will be reviewed. Next, the use of students as standardized patients will be outlined and finally, the impact of participation on SPs will be explored.

2.2 Performance-based Assessment

Educational measurement has been enhanced in recent years through performance-based assessment (PBA). This was first described as evaluation methods that require “the examinee to demonstrate specific skills and competencies... to apply the skills and knowledge that have mastered...the examinee’s task is to construct an original response, which the examiner observes and evaluates” (Stiggins 1987) Forms of PBA include four
basic components: a reason for assessment, a specific performance to be evaluated, exercises which elicit that performance, and systematic rating procedures.

The popularity and acceptance of problem-based assessments in the arena of public education stems from their ability to measure the higher order critical thinking skills such as analysis, synthesis and evaluation, described in Bloom's taxonomy (Bloom et al, 1971; Masters et al, 2001), and emphasize proficiency in processing skills: problem solving, comprehension, reasoning and meta-cognitive processes. (Linn et al, 1991) This is represented schematically by this author in Figures 1 and 2.

Some believe that evaluation of complex cognitive tasks make PBAs a multidimensional measurement. (Swanson et al, 1995, Traub, 1994) Traub notes the interaction of candidates in performance assessment tasks in numerous generalizability studies indicates the measured construct is multidimensional.

Psychometric and educational literature debates issues of reliability and validity of scores obtained through PBA. Some educators insist that performance (PBA) measures be held to the same stringent standards of reliability and validity as standardized, norm-referenced assessment (Burger and Burger, 1994): others argue that reliability has too often
been overemphasized at the expense of validity while validity itself has been viewed too narrowly. (Linn et al, 1991) Linn and colleagues support the acceptance of lower reliability coefficients for PBA compared to traditional format MCQ examinations. Although the multidimensionality of PBAs will result in decreased internal consistency of scores across task items, it enhances the relevance and validity of the test. They feel while reliability as an important issue should not be ignored in new assessment forms, it should not be the only or primary criteria in judging the quality of the measurement. They call for an expanded approach to validation of PBAs that embrace concepts beyond test validity such as consequences of testing, test fairness, content coverage and quality, generalizability, cognitive complexity and meaningfulness.

These views are supported by Messick in interpreting Cronbach’s basic tenet to understand validity as not a property of the test but of the meaning or interpretations made on the basis of test scores. (Messick, 1995)

To improve the validity and reliability of oral examinations, researchers suggest increasing tasks and evaluators in PBA to decrease the impact of specific tasks on variance components which involve tasks interactions (eg. tasks by candidate, tasks by rater) (Leichner et. 1986; Traub, 1994) For this reason, the objective structured clinical examination (OSCE) has become the preferred form of PBA used in clinical competency assessment as it typically tests greater numbers of tasks, using greater numbers of examiners, thereby minimizing task and examiner related variance components.
2.3 The Objective Structured Clinical Exam (OSCE)

The objective structured clinical examination (OSCE), first described by Harden et al (1975) has become an accepted format for evaluating clinical performance in the health care arena, as it is considered one of the most valid and reliable approaches, especially for a large group of candidates. (Harden, 1979) It is widely used in undergraduate curricula and for high stakes licensing and certification bodies in North America. (Van der Vleuten & Swanson, 1990)

While OSCE structure may be varied, three key components are present: a highly structured format of task stations, objective scoring systems, and use of standardized patients (SPs) to portray the clinical problems in a consistent manner. Generally candidates must rotate through a series of timed stations chosen to reflect important components of clinical competence. The stations may require the examiner to use a checklist, which records behaviour. Alternatively, global rating scales are used alone or with checklists to interpret clinical competence.

SP-based OSCEs have been used to assess a broad range of clinical skills. Most often, they are used to measure history taking, physical examination, and communication skills, although skills in diagnosis, laboratory utilization and patient management are sometimes tested through oral exams or written questions as a follow-up to an SP encounter. Candidates spend between five and thirty minutes, depending on the exam at each station, which may vary in number from 3 to 20. (Van der Vleuten & Swanson, 1990) In general, the length of the station is determined from the perspective of content validity. It has been demonstrated
that in a clinical assessment OSCE, candidates did not, on average, take the full time interval of 15 minutes allotted. (Chambers et al, 2000)

The medical education literature contains numerous studies that describe the use of OSCEs as measures of clinical competence of undergraduate medical trainees in general, as well as in a number of specialty domains, including psychiatry (Hodges et al, 1997, Hodges et al, 1998), surgery (Cohen et al, 1990, Mann et al, 1990), pediatrics (Hilliard and Tallet, 1998) and physiatry (Jain et al, 1997) In addition, licensure OSCE examinations for Canadian-trained physicians (Reznick et al, 1992, Reznick et al, 1993) and for family physicians in Quebec (Grand'Maison et al, 1992) have been described. OSCEs have also been used to evaluate competency in addressing medical ethical issues (Cohen et al, 1991) and to select candidates for an Ontario-based training program for foreign medical graduates. (Cohen et al, 1991)

In the extensive study of OSCEs used in these disciplines, they have been established as PBA instruments with good validity and reliability. (Kolovitz et al, 1991; Matsell et al; 1991. Sloan et al, 1993; McFaul et al, 1993; Cohen et al, 1988) Since OSCEs have been shown to have much higher reliability and validity than traditional and less structured oral examinations, they have grown in acceptance as a means of episodic PBA. (Rothman & Cohen, 1995)

2.4 Assessment Tools: Checklists and Global Rating Scales

During an OSCE, the candidate is assessed by the examiner in a standardized fashion using binary content checklists for knowledge variables and/or global rating scales for
process variables. Usually, the examiner ticks off checklist items while the candidate interacts with the standardized patient, and in some cases, the patient will complete the checklist following the interaction. In addition, global ratings may be competed by examiner or standardized patient following the encounter. Content checklists are commonly written by qualified practitioners or examiners who list a series of items the candidate should correctly address, and in most OSCEs, may account for the bulk of marks awarded to candidates. Global rating scales are typically designed on a 5 point scale which may rate the candidate in separate domains such as empathy, organization and focus, verbal and non-verbal skills, or overall impression of the interview.

The 5-point scale can be interpreted in terms of the five levels of learning described by Collis and Biggs as the Structure of the Observed Learning Outcome (SOLO) taxonomy. (Collis & Biggs, 1979). This scheme is well-designed to reflect the quality of learning of a complex set of skills such as process analysis and discriminates between performance levels described as pre-structural, uni-structural, multi-structural, relational and extended abstract (master or insightful). This author has represented the hierarchy schematically in Figure 3. Dreyfus and Dreyfus similarly propose that professionals pass through five stages in the development of expertise: novice, advanced beginner, competence, proficiency and expertise and that a distinct form of problem solving
characterizes each level in the developmental process. (Dreyfus and Dreyfus, 1986) Hodges and colleagues elaborate this theory to explain that the early novice stage is characterized by the collection of large amounts of data in no particular order and with little regard for situational factors, which are then used to synthesize a problem solution. An expert gathers much more focused information, relies on many different types of data including situational variables, compares all observations with previous experiences and quickly responds without a formal process of problem solving. (Hodges et al, 1999)

Schmidt, Norman and Boshuizen use a four stage model to describe development of medical expertise: (1) development of elaborated causal networks, (2) compilation of elaborated networks into abridged networks; (3) the emergence of illness scripts; and (4) the storage of patient encounters as instance scripts. Clinicians encapsulate large amounts of knowledge in order to interpret large amounts of data expeditiously and without effort, whereas more effort would be required for the effort to deconstruct this material into the formal thoroughness of the novice. (Schmidt, et al, 1990)

Support for this theory is demonstrated in a behavioral study of clinicians in Leeds showed that senior clinicians had a more flexible approach to interviewing, asking fewer yet more relevant and revealing questions compared to stereotypical interviewing elicited by junior clinicians. (Leaper et al, 1973)

The evidence suggest that while checklists used in OSCEs may reflect the approach of novices to clinical problems, they may not detect the complex, hierarchical problem-solving of expertise which processes by simultaneously gathering and integrating information. Checklists may, in fact, penalize the mastery of efficiency at the expense of attention to detail. (Hodges et al, 1999)
Hodges and colleagues suggest that it may be meaningless to speak about the validity of an OSCE without specifying the type of measurement instrument used as content checklist or global rating, and one must consider that an instrument that is valid at one level of training may not be valid at another. (Hodges et al, 1998) They showed evidence of construct validity in terms of significantly better scores with psychiatry residents vs clinical clerks in the case of global ratings, but no difference in checklist scores between clerks and residents. In another study, scores on global ratings were higher for experts while scores on binary checklists declined with increasing levels of expertise. Global ratings were useful in discriminating clerks from family physicians, but not in discriminating residents from clerks, proving more sensitive than checklists yet not sufficient for measuring increasing levels of expertise. (Hodges et al, 1999) Global ratings may better reflect accuracy whereas checklists reward thoroughness, rather than competence (Norman, 1993) and may penalize efficiency. (Norman et al 1985)

While global rating scales have been shown to be valid and reliable used with checklists (Cohen et al, 1991; Martin et al, 1997; Reznick et al 1998) or alone (MacRae et al, 1995) they may be more reliable across stations (Cohen et al, 1996, Regehr et al, 1998; Regehr et al, 1999) show better construct validity (Hodges et al, 1998; Regehr et al, 1998), demonstrate more reasonable generalizability across stations and across examinations (Regehr et al, 1999), and have better concurrent validity (Regehr et al, 1998) The binary checklists typically used in most OSCEs also neglect higher components of clinical competence in process variables such as empathy, rapport and ethics, which can be captured by global ratings (Cox, 1990) and which may increase with years of training. (Hodges et al. 1998)
2.5 Use of Standardized Patients

For patient stations, either actual patients with stable physical findings, or standardized patients have been used. (Davis Feikert, et al, 1992) The standardized patient (SP), first proposed by Barrows & Abrahamson (1964) is a normal individual who is taught to simulate every aspect of a patient's illness in a totally consistent manner so accurately that the simulation cannot be detected by a skilled clinician. (Barrows, 1971) Often, professional actors are used (Barrows & Abrahamson, 1964; Lamont & Hennen, 1972; Meadow & Hewit, 1972; Davis et al, 1985; O'Hagan, Davis & Pears, 1986). They can lack symptoms, have abnormal findings on physical exam, or simulate physical findings (eg diminished breath sounds, cough, etc.) Students interact with standardized patients as though they were interviewing, examining, and counselling real patients. Often, standardized patients are trained to complete checklists and rating forms at the end of encounters, recording the details of the encounter regarding information obtained, examinations performed, and counselling provided, as well as rating communication skills of examinees. Some writers have attempted to identify the personal characteristics that SPs should possess to meet the demands of their work. (Barrows, 1993; Norman et al 1985; Stillman et al, 1980; Tamblyn et al, 1990; Davis, 1989; Whitehouse et al, 1984) Important considerations in hiring SPs include such characteristics as "intelligence, appropriate motivation, flexible work schedules, stamina, and emotional and geographic stability. (King et al, 1994)

Since they were first introduced by Barrows twenty years ago, ever-increasing numbers of standardized patients have been incorporated into the instructional and evaluation programs of medical and nursing schools, as well as certifying and licensing bodies.
(Barrows, 1971; Holmes et al, 1978; Barrows et al, 1982) The use of SPs is currently widespread. A recent survey found that 80% of American and Canadian medical schools employ SPs for teaching and assessment purposes. In 1994, 20 US schools required students to pass an SP examination to graduate, Canadian medical schools use a national SP examination, and the National Board of Medical Examiners (NBME) will implement a similar exam in the US in 2001-2. (Klass, 1998) In Canadian pharmacy, SPs in OSCE formats are used in many programs, including the Ontario College of Pharmacists in the quality assurance re-certification process (since 1998), the Pharmacy Examination Board of Canada (beginning May 2001) and in two undergraduate courses at the Faculty of Pharmacy, University of Toronto for formative and summative assessments (Sibbald, 1998; Austin, 2000). In North America, the work of SPs is a recognized occupation. (McNaughton et al, 1999)

It is important to distinguish between the many different roles for people who are used for teaching and assessment in medical education, as not all of them are associated with standardized patients. For example, 'role-playing' is a technique in which the student takes on the role of patient or health care professional. The educational advantage is derived by the role player, who gains first-hand experience about what it is like to be in the role of the dependent patient or responsible clinician. 'Practical instructors' are clinicians who assume the role of the patient in order to teach the student demonstrable techniques: for example, the proper use of devices such as medication inhalers, glucose meters, or blood pressure cuffs to pharmacy students, or the techniques of pelvic or genitor-rectal examinations to medical student. This assists the student in developing the skills and sensitivity they need to perform or demonstrate these functions. 'Patient instructors' are actual patients who have been
carefully educated about their own illnesses and how they should be evaluated. They can give feedback to the student about the accuracy and completeness of an interview. They see themselves as experts in and as exemplars of the condition and contributors to students' education and training. (Spencer et al, 2000) Geoff Norman has coined the term 'standardized' patient to replace the term 'simulated' patient because it underlines the major advantage of this technique: to provide a patient problem that will not vary from student to student. Barrows refines this definition to include either a simulated patient, (actor) or an actual patient carefully coached to present their own illness in a standardized, unvarying way. Thus the term standardized patient will not reveal to the students whether they are actually dealing with an actor or a real patient in their teaching and assessment contacts. (Barrows 1993)

Patients selected to be SPs conform to standard criteria. They need to be very intelligent, available, geographically stable, flexible, interested in medical education and improving the quality of health care delivery, and have good communication and teaching skills (Stillman, 1993)

2.6 Value of SPs for Teaching and Evaluation of Clinical Skills

Enthusiasm for the use of standardized patients is generated by consideration of their demonstrated value. There is a range of difference in the use of standardized patients. They may be trained to portray a patient, who actually exists, with the complexities of real circumstance, or they can portray totally fictional cases in which history and findings are simplified for teaching purposes. The most elemental application of standardized patients is
for teaching and formative evaluation in learning clinical skills. The next most useful application is for skills assessment at the end of a course or program, and the third most useful application is to evaluate competency before graduation or as a component of certifying or licensing examinations. (Stillman, 1993) Analysis of the worth of standardized patients can be summarized with respect to four areas: value added to students, value added to teaching, value added to assessment and advantages to educational programs.

2.6.1 Value of SPs to Students

"For the junior student in medicine and surgery it is a safe rule to have no teaching without a patient for a text, and the best teaching is that taught by the patient himself" (Osler, 1904)

Clearly, standardized patients who are simulating an illness offer advantages to the student over real patients when used in education settings. Students learn to cope with actual clinical problems in a less threatening environment than when working with ill patients on a hospital ward or in a clinic, where students may be overwhelmed with multidimensional problems. (Van Daken et al., 1989) Pharmacy students who deal with ambulatory patients in a community store often have difficulties with both the urgency of such patients’ demands and the responsibility of assisting patients who will not consult with any other health care worker. Standardized patients provide a transition to the real patient: students can learn and practice formative skills in a setting free of awkwardness that safeguards patients and provides immediate, consistent, and corrective feedback on their performances. Students can practice without embarrassment about their novice status as they learn to gather information from the standardized patient, and they can perfect their techniques until they become confident. The
standardized patient thus enhances the value of the real patient, from whom students then can learn without concerns about their ability to communicate. Patients also will feel as though they are being cared for by competent professionals rather than neophytes. Researchers have shown that students demonstrate an increase not just in skill but in confidence in giving information, handling emotions, and motivating patients, through practice with standardized patients, while becoming more critical in self-assessment. (Hoppe et al, 1988; Thomas, 1992)

Standardized patients can be used in a variety of techniques to enhance formative learning. Using standardized patients allows the interruption and restarting of the encounter to permit student feedback and practice. Unlike real patients, the standardized patient can pretend lack of awareness of a 'time out' discussion. The students and instructor can debate communication and therapeutic issues, shaping the students' thinking and then allowing them to continue with the patient as if nothing had happened. The encounter can be started over again, if necessary. The passage of time can be ignored, and the patient can be encountered at various time intervals, allowing the student to learn continuity of practice at one setting. Students can encounter the same patient at the beginning and at the end of their course experience, to assess improvement in skills. Reflective learning, especially if assessed against a set of explicit criteria, can improve learning. The experience of Usherwood has demonstrated an increase in effective, sensitive interpersonal communication skills, through subjective and behavioural evaluation. Eighty-five percent of students felt their interview skills had been improved through use of standardized patients, and many students commented on other learning enhancements, such as confidence, self-awareness, structure, and alternative verbal behaviours. These results showed statistically significant increases in number of open questions, and questions referring to feelings, beliefs or behaviour.
This type of formative practice is critical to the success of the standardized patient used for summative examination assessment.

Standardized patients make it possible for students to practice communication skills with difficult patients. Wolfe has observed that students learn how to handle emotionally-charged situations in such controlled settings, responding with integrity, empathy, compassion and respect. (Wolfe et al, 1987) Campbell has demonstrated improvement in HIV/AIDS specific interaction skills in trained students compared with students who did not train with standardized patients. (Campbell et al, 1996) Important changes in physicians' communications skills have been reported after a program with simulated patients, showing that trainees used significantly more problem-defining and emotion-handling skills without increasing visit length, reported more psychosocial problems, engaged in more strategies for managing emotional problems with actual patients, and scored higher in clinical proficiency with simulated patients, compared with untrained physicians. Patients of trained physicians reported reduction in emotional distress for as long as 6 months. (Roter et al, 1995) Gettman's study using simulated patients with pharmacy students showed that women may be more adept in communicating feelings and emotions and that same-sex counselling is more comfortable than opposite-sex counselling. Gender differences in patient interviewing may be improved with increasing use of focussed techniques during simulated patient training. (Gettman et al, 1996) Standardized patients focus students' attention on their performances as practitioners, motivate them, and encourage active learning. Students learn specific skills required during patient interviews, such as clinical details of diagnosis, appropriate history-taking skills, management issues, attitudes and continuing care. (McAVoy, 1988)
2.6.2 Value of SPs to Teaching

"The concept of teaching is the deliberate creation of situations from which learners cannot escape without learning." (Cowan, 1988)

Apart from the student perspective, the use of standardized patients has significant value to teaching methodology. In order for SPs to teach, what they should teach should be clearly defined, and be reinforced with training so they can function as extensions of faculty. While they cannot teach medical content, they are wonderful teachers of process, interviewing skills, patient education and counselling skills, and can reinforce on feedback with students which checklist items were obtained (Stillman, 1993)

Standardized patients solve ethical and practical teaching problems. Their use avoids the ethical problems which may emerge when authentic patients are interviewed outside the confidential atmosphere of the health-care professional-patient relationship. (Kinensley, 1993) Situations where ethical concerns exist for the comfort of real patients can be solved by using actors, for example in the teaching of genito-urinary examination skills for medical students. (Hawkins et al, 1997)

The acceptance of 'simulators' as participants in clinical teaching is encouraged by the report, *Physicians for the 21st Century*, which recommends that students should be engaged as "active learners" rather than "passive recipients of information". (GPEP report, 1984) Standardized patients give students a consistent learning experience when there are constraints of student group size, time and scheduling within a traditional curriculum and offer the opportunity to make assessments in ambulatory settings, which is often difficult
with real patients. (Neiman & Thomas, 1987; Furman et al, 1994) Standardized patients make it possible to provide students with the same or equivalent patient experiences instead of the random experiences of cases presenting on the wards and in clinics. (Mast, 1984) Thomson has shown significant differences in student performance between cases, not reducible by manipulation of the scoring method. Reducing as much of the variance due to context as is possible with standardized patients increases the reliability of the assessment. (Thomson, 1990; Van der Vleuten & Swanson, 1990)

Standardized patients can be used to develop a core set of problems and experiences wherein faculty have a chance to plan and control the learning process. Apart from systematic planning of content complexity of therapeutic problems, attention can paid to communication variables. Children (Woodward & McConvey, 1995) and teenagers (Hardoff & Schonmann, 2001) have been used as standardized patients to focus attention on special communication techniques for the young and the adolescent. It is important that case structuring also address perception variables. It has been demonstrated that student judgements may be unrelated to clinical skills, but due to a bias regarding the patient, such as age, sex, ethnicity, accent and physical appearance, as well as perceptions regarding patient’s competence or social standing. (Gerbert, 1984; Johnson et al, 1986) It is possible to train standardized patients who are systematically varied to heighten the awareness of the student with respect to ingrained stereotypes that likely influence patient bias and form barriers to accurate clinical assessment.

Use of standardized patients encourages faculty to determine what should be taught at what stage during the educational program, using the simulations as a means to ensure that the agreed-upon themes and issues are covered. It draws the attention of faculty to learning
objectives and methods of assessing their achievement. Faculty can control the content, domain, sampling emphasis, and level of complexity of clinical problems by using standardized patients in specific scenarios.

Use of standardized patients requires less faculty involvement in time intensive tasks such as directly teaching history taking, and in locating appropriate patients for teaching. At the same time, it raises faculty involvement in education, forcing a critical look at the curriculum and catalyzing collaboration with other health care professionals.

2.6.3 Value of SPs to Assessment

In addition to the teaching perspective, the use of standardized patients has significant value specific to assessment methods. Standardized patients provide the opportunity for faculty to observe and judge students' performances in gathering information or providing counselling. Faculty can directly assess how students incorporate knowledge and skills into practice, rather than attempting to infer clinical performance from case presentations and medical records. In fact, the use of standardized patients for history-taking and communication skills are consistently more reproducible than scores for other assessments of clinical skills, and research indicates it is more efficient to use standardized patient tests for these skills, and give written tests for other components of competency. (Stillman et al, 1986; Stillman et al, 1986; Swanson, 1987)

Standardized patients themselves can give feedback to both students and faculty about communication techniques, providing the patient's perspective in the assessment of interpersonal skills. It has been shown that standardized patients are accurate and consistent in assessment, and maintain their accuracy in recording and assessing throughout a day without the impact of fatigue, due to training and case knowledge. (Vu et al, 1992; Wang et
al, 1996) Standardized patients in particular, can assess the use of empathy, an enabling factor in clinical competence. After competence, empathy may be the characteristic most valued by a patient, and since most patients lack the expertise to determine competence, empathy may be the key assessable component. An analysis of the assessment of empathy by standardized patients showed that trained actors rated empathy more objectively than real patients: they have received training in assessing, and the focus of their attention is the accurate evaluation of student performance, not the medical problem, as in the case of a real patient. Actors themselves provide a valuable means for student evaluation and feedback. (Colliver et al, 1998)

Standardized patients provide consistency and control of assessment by assuring a predetermined and repeatable setting. They are a method to regulate the clinical problems used in performance evaluation. Curricular objectives can be matched with assessment methods to directly measure achievement. The faculty can control the criteria used to judge the adequacy of performance by constructing specific clinical problems against which the performances of all students can be compared, using standardized patients to enable more objective assessment, minimizing biases and inconsistencies inherent in the use of real patients. With proper controls, it has been shown that groups of examinees at more advanced levels of training perform better on standardized patient testing than do students at earlier levels, supporting the validity of this measure. (Barnhart et al, 1995) The level of complexity evaluated can be increased systematically to correspond with appropriate milestone levels of student ability.

The use of standardized patients requires faculty to define performance criteria and standardize the teaching of clinical skills to students. The assessment of clinical skills with
standardized patients is an important balance to the measurement of knowledge and judgment by written tests. Like the multiple choice test, it is standardized, can be given in multiples, can be scored in reliable and valid ways, yet it provides a much more valid assessment of clinical competence and curriculum effectiveness than written tests. (Barrow & Peters, 1984; Barrows et al, 1987; Barrows, 1993) Written tests focus on more purely cognitive skills, which are different components of competence, and only indirectly assess clinical behaviour. The standardized patient provides authentic assessment that directly examines the behaviours that are important in a competent clinical performance.

2.6.4 Value of SPs to Educational Programs

Lastly, the use of standardized patients provides advantages to educational programs. In 1984, the AAMC’s report on the General Professional Education of the Physician (GPEP) criticized the lack of a sound system for evaluating students in the clerkships and recommended that medical faculties develop procedures and adopt explicit criteria for the systematic evaluation of students’ clinical performance. (Muller, 1984) In 1986, the AAMC sponsored a national conference on the evaluation of students’ clinical abilities, establishing a special interest group on standardized patients. A consensus conference on the use of standardized patients in the teaching and evaluation of clinical skills was held in 1992, in Washington. The proceedings were published to serve as a catalyst for continued research, stressing the value and role in education and evaluations, and providing a template for the useful application of these patients in clinical skills teaching and assessment, to promote broader use. Standardized patient use has prompted discussion between educational administrators, researchers and faculty to review teaching techniques and outcomes and explore their use not just in undergraduate programs, but in high-stakes applications such as
in certifying or licensing examinations. Medical school programs have prompted change in other health care specialities: the Ontario College of Pharmacists spearheaded a practice review program in 1997, utilizing standardized patients as one measure of performance.

There is a high potential for use of standardized patients in continuing medical education programs (Davis et al, 1992) as well as blinded observations of physicians’ performance studies. Groups from McMaster in Hamilton and from McGill in Montreal have conducted studies using standardized patients for blinded assessment of quality of care of physicians during clinic visits. (Norman et al, 1985; Tamblyn et al, 1997)

Once in place, standardized patient programs can be structured to provide feedback to the faculty about program and teaching effectiveness. Newble, in Adelaide, has emphasized the importance of considering the impact of standardized patient testing on education in general. To develop a program, faculty need to reach consensus on what should be learned, and performance criteria must be concretely defined. This process leads to more standardized instructional experiences and learning outcomes, particularly if carefully communicated to both faculty and students. (Newble & Jaegaer, 1983; Newble, 1988)

Standardized patient programs can stimulate collaboration between institutions for teaching and assessment methods that will foster uniformity in educational approaches. Because of the expense in faculty time and effort required to develop and implement effective clinical performance assessments and the need to market-test and disseminate these methods, six consortia of medical schools have united to facilitate the best use of limited resources available: North Carolina, Gulf Coast, Northwest, Southern California, Metropolitan New York, and Upstate New York. Although all consortia share similarities, each has its own vision, purpose and operational methods. The benefits identified and
acknowledged include sharing workload, access to consultation and training expertise of experienced case writers, standardized patient trainers, educational researchers and psychometricians and educational administrators. Sharing of physical equipment, testing facilities, and patients was also feasible for some consortia. Research possibilities expand greatly with the enhanced numbers involved in multi-school projects. The most significant advantage may be the intellectual strength and support that consortium membership affords all participants: group strength, internal and external credibility. (Morrison et al, 1994)

This kind of collaboration has also been expanded globally. The Educational Commission for Foreign Medical Graduates (ECFMG) Clinical Competence Assessment (CCA) program is being introduced in pilot projects worldwide, after feasibility studies in fifty countries. Standardized patient cases will be developed and then modified with medical faculty of each country to establish a regional context. It is hoped that these pilot projects will result in an assessment that will be equivalent across cultures and national borders, so that global standards can be established for the use of many nations to meet their individual and collective needs for world-wide health care.(Sutnick et al, 1994)

2.7 Technical Issues of the Logistics of Using SPs

"Just Do It" (Miller)

or.... ?

"Doubt is the Road to Truth" (the Koran)

A consensus conference on standardized patients produced a summary of technical issues to consider when using SPs (Stillman, 1993). They emphasized the importance of beginning to use SPs in introductory courses. Faculty support and involvement are critical to
success, and their involvement in hands-on experience, research and publications, design in teaching, support groups for sharing information about SPs, workshops and standard settings is essential. Templates should be used in case development and cases are better developed by generalists than specialists.

While fatigue over the course of an examining session may potentially introduce systematic or non-systematic error in assessment of communications skills, no evidence was found of this effect in a study which compared examiner and SP-raters (Humphries & Kaney, 2001)

Training standards should be established with trainers working with faculty and having debriefing sessions with SPs. The SP should have the ability to portray the case authentically and also good observational and recall skills. Time of training ranges from one to two hours per case to ten to fourteen hours depending upon complexity of role and recording and evaluating functions: it is easier to train role-portrayal than use of rating scales. (Stillman, 1993)

Costs and resources are a major determinants is the use of SPs. The amounts paid SPs range from $7.50 to $50.00 per hour (US) and the cost per student for an exam ranges from $150.00 to $200.00 per student, and include cost of SPs, a trainer, faculty development, and related equipment. (Stillman, 1993)

2.8 Fidelity of Assessment with SPs: Issues of Realism, Reliability and Validity

"Any psychometric assessment of any specific method or format (eg OSCE) applies only to the interpretation of each administration of the format or method. Reliability or validity are not inherent qualities of the method or format. They must
be determined for each instance of use." (Dauphinee & Blackmore, 2001)

Statistical tests of OSCE results are undertaken in order to draw inferences about the ability of the candidate that extend beyond the sample items to a larger domain. In order that inferences are reproducible (reliable) and accurate (valid), test design should develop a sampling plan that reflects the skills and areas to be assessed. Extensive literature has reviewed issues of realism, reliability and validity with the use of standardized patients. (Swartz et al, 1996) The psychometric properties of the examination are well established, with the SP-based OSCE regularly demonstrating reasonable internal consistency, inter-rater reliability, construct validity, and concurrent validity in a variety of skills. (Hodges et al, 1998; Prislin et al, 1998; Jain et al, 1997; Mohtadi et al, 1995; Singer et al, 1996) This has been documented for the traditional binary checklists which are station-specific as well as for the five-point global rating scales which are station-independent. (Rothman et al, 1996; Cunningham et al 1996; Cohen et al, 1996; Regehr et al 1998)

Both realism and reliability have been well documented in testing both communication and biomedical skills (Tamblyn et al 1991; Schnabl et al 1991; Vu et al, 1992; Colliver et al, 1993; Hodges et al 1996; de Champlain et al 1997; de Champlain et al 1998; Hodges et al 1997; Boulet et al 1998).

Realism: Research has shown that experienced physicians cannot differentiate SPs from real patients in their offices. (Proceedings AAMC, 1993; Annex to Proceedings AAMC, 1994) and medical students are poorly able to discriminate SPs from real patients (Sanson-Fisher et al 1980) A qualitative survey showed 93% of psychiatry residents who participated
in an OSCE rated the scenarios very realistic and 80% considered the scenarios reflective of clinical situations they had experiences. They were very confident that an OSCE could discriminate between incompetent and competent medical students but less certain of its ability to do so at the resident level. (Hodges et al. 1999)

**Reliability:** The reliability of scores using SPs, reported as 0.69, may be increased to recommended levels by increasing the number of cases, while the reliability of pass/fail decisions, often of more concern, is better, for example 0.96. (Colliver et al, 1989; Colliver et al, 1997) It has also been suggested that researchers may be striving to reach a reliability score that is too high, and that 0.70 may be more realistic than 0.80, with standardized patients (Stillman, 1993) Reliability is less important in applications which teach and judge beginning competence and provide formative assessment that reinforce education, but more important when a student's academic progress or graduation is to be decided. (Stillman, 1993) Further, while inter-station reliability can be improved by using more stations, competence is specific to content in some degree and not an entirely generalizable measure. (Vu and Barrows, 1994) While OSCEs should demonstrate some degree of inter-station reliability, more importantly, they should be designed to ensure that the clinical cases presented sampled represent as wide a range as possible of situations a competent practitioner would be expected to handle. (Colliver et al, 1989)

Reliability may be maintained throughout an extended testing period with a number of SPs simulating the same case (Battles et al 1994; Gomez et al 1997; Colliver et al 1998). Accuracy in both assessment and portrayal can be maintained in different sites (de Champlain et al, 1999) or with multiple SPs representing the same case. (van der Vleuten et al, 1990; de Champlain et al 1998; Doig et al 2000) Multiple SPs have little or no effect on
examination reliability (Swanson & Norcini, 1989; Colliver et al, 1990; Colliver et al, 1993) For example, in one study of patient satisfaction ratings with multiple SPs, the generalizability coefficient free of multiple SP measurement error was 0.70, and the coefficient that reflects measurement error due to multiple SPs was 0.74 (Colliver et al, 1993).

The degree to which OSCE measures generalize across multiple administrations to the same students, that is, stability over time and across content domains has also been described by Regehr and colleagues. Their analysis reported low generalizability for checklist scores both across stations within an examination and across examinations from different domains. In contrast, the global rating scores demonstrated reasonable generalizability across stations and across examinations although specific items on the scales varied for different exams. Since these data supported the superiority of global rating scales in assessing a broader set of process skills as opposed to content-specific checklists, the authors suggested that global rating scales be increased in weighting and reported separately from checklist scores. The consistency in global rating scores over time may provide a good tool for identifying students who need remediation and thus a uniform set of global scales should be recommended for use among coordinated courses. (Regehr et al, 1999)

"Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few." (Ebel, 1961)

*Validity:* Validity is the accuracy of a prediction or inference made from a test score (Cronbach, 1971). It is not a property of the test itself, but of interpretation of test scores. It is through validation studies that the meanings of test scores are established. The process of establishing validity is gradual and cumulative, involving many studies in three areas: (1)
content validity – logically analyzing the content of the examination, (2) concurrent validation - study of the relationship between scores and independent measures of competence. (Colliver & Williams, 1993) and (3) construct validation – study of differences in group performance (eg) establishing relationships between candidate performances on the test and other characteristics of candidates implied by the construct (eg clinical competence), such as level of training. Research on validity is less well established since a generally accepted indicator of clinical competence is needed for use as a criterion standard and to guide scoring and standard-setting. (Pieters et al 1994; Colliver, et al 1997)

Content validation: Content validation should be the major concern in achievement testing (Kane, 1982; Frederiksen, 1984) and SP-based tests seem amenable to direct validation. It should be easily built into test design by careful construction and sampling. However, studies are needed to address variables affecting accurateness such as effect of the type of scoring (checklist vs global rating), lack of training of candidates causing misconceptions about testing format, time constraints, and bias of observer raters. (van der Vleuten & Swanson, 1990)

Concurrent validation: Concurrent validations studies have shown a lack of correlation between scores with the OSCE assessment method and academic measurement that mainly reflect knowledge such as written examinations (Norcini 1994; Gomez et al 1997; Rosebraugh et al 1997) or MCAT or NBME scores (Croen et al, 1994)

Construct validation: A number of studies have contributed to construct validation efforts by showing that examination performance is higher for examinee groups with more training: second-year vs fourth year medical students (Barnhart et al, 1992); medical students vs residents (Petrusa et al 1987; Newble et al, 1981); and residents at different levels of
training (Stillman et al, 1986; Stillman et al, 1991; Stillman et al, 1982; Robb & Rothman, 1985) One study showed third and fourth year internal medicine residents performed one-half to eight-tenths better than did first-year residents, with second-year residents falling between these two groups (Stillman, 1991) while two others showed little increase in third year medical students performances on an SP examination as they progressed through clinical clerkships (Petrusa et al, 1987) and deterioration in communication skills of third year medical students over one year instruction. (Prislin et al, 2000) However, construct validation results may be largely dependent on the assessment tool used.

Checklists have been shown to be a valid assessment for clinical clerks, but not for residents or physicians in practice, while global rating scales, more sensitive to levels of expertise are better in that they can discriminate clerks from family physicians, but not residents from clerks. (Hodges et al, 1998; Hodges et al, 1999) Since these assessment instruments appear to vary in their validity depending on the level of expertise they are testing, these authors suggested a need to explore further ways of differentiating competence at various levels in PBA by designing instruments which focus on the candidate’s sequencing and types of questions, and the degree to which they reflect information processing.

2.9 Test Security

A review of technical issues in the logistics of using SPs concluded that there is little problem with test security, especially when the SP application is measuring process rather than the ability to recall content and produce answers. Security should be monitored, but the
issues should not restrict use. (Stillman, 1993) Five studies based on data for at least ten examination forms administered at 23 sites have shown no consistent systematic increasing (or decreasing) trend in scores throughout the examination period, as would be expected if there were widespread violations of test security. (Rutala et al, 1991; Coliver et al, 1991; Williams et al, 1992; Stillman et al, 1991; Colliver et al, 1992) Previous research has indicated that the global rating performance of interpersonal skills are most resistant to breaches of security, with scores varying by 2.15%, while severe disclosure of checklist items to 50% of candidates who knew they would be seeing the same cases, led to performance gains of 7% versus 2% variation for non-disclosed cases. (De Champlain et al, 1999)

2.10 Comparison of SPs vs Faculty as Raters

An important consideration in test design is the determination of who should rate candidate performance. While previous work suggested physicians (ie faculty) raters are naturally either stringent (hawks) or lenient (doves) (Newble et al, 1980; Lubbrook & Marshall, 1971) more recent work shows differences in accuracy among these raters were almost eliminated by rater training. (van der Vleuten et al, 1989) Several studies have compared the evaluation of skills by faculty vs SPs. Adequate inter-rater agreement can be achieved through use of either SPs or faculty raters. SPs have been found to give more positive evaluations of communications skills of undergraduate medical students than faculty, reflecting differences in perception in the role of observer and patient, and proposing
that the patient, as the ultimate participant, is a more appropriate assessor of interpersonal skills. (Cooper et al. 1998) Authors have reported that the higher ratings of SPs vs faculty might be due to errors of commission and to multitasking (patients and raters) with fatigue playing a role. (Cushing et al. 1996) Other studies have reported either poor correlation between scores given by medically trained faculty and ‘consumers’ (SPs) (Thomson 1993; Finlay et al. 1995) or high-level agreement (Kopp et al. 1995) A test of the reliability and validity of faculty raters showed low correlations between raters and between performances as compared to experts. (Kalet et al. 1992) In a study of ratings of a medical interview, faculty ratings were higher than SP ratings, which were higher than student self-ratings: faculty ratings generally did not correlate with the others. (Furman et al. 1996) For history taking tests, physician raters have been shown to differ less from gold standards than SPs rating from recall, but less significantly than SPs rating during observation. (Martin et al. 1996) In a comparison of evaluation methods, SP raters showed similar patterns to physician raters in demonstrating equal or greater reliability of global ratings to checklists. (Regehr et al 1999)

The most important factors in rater selection may be practical and educational considerations. It is desirable to have faculty as raters since the observation of candidates provides useful feedback concerning instructional effectiveness in terms of the skill levels of trainees. (van der Vleuten & Swanson, 1990) while SPs as raters capture performance levels from the patient’s perspective of communication needs, such as establishing rapport, sensitivity to patient needs, attitudes and avoidance of jargon.

SPs have also demonstrated equal or greater effectiveness in teaching interpersonal skills compared to faculty (Vanatta et al. 1996; Kleinman et al. 1996; Sachdeva et al. 1997)
2.11 Use of Students as Standardized Patients

"When one teaches, two learn." (Author Unknown)

Using professional actors as SPs is costly and labour intensive: authors report a time investment ranging from four (Barrows, 1971; Baerheim et al, 1995) to 10-25 hours (Helfer, Black & Teitelbaum, 1975; King et al, 1998; Yelland 1998; de Champlain et al 1998) to train a standardized patient. The use of students as SPs is an alternative method to outside professionals. The consensus group on technical issues associated with SPs suggested use of senior students as SPs is cost effective and affords them an important understanding of what it is like to be a patient (Stillman, 1993). There is little experience reported in the literature.

The first study was the use of senior medical students as patients and preceptors to introduce basic history and physical examination skills to second year medical students for formative instruction in a clinical medicine course. (Barnes et al, 1977) Seniors were equally effective as resident, fellow or faculty staff in teaching techniques and statistically better than staff in providing the sophomore with both suggestions as to how to improve their techniques and how their approach might affect the patient's attitudes and behaviour. Seniors felt they learned more about the techniques and the patient's viewpoint. The authors concluded seniors could better remember difficulties in communicating with patients, and felt more at ease relating to students than staff, while providing immediate feedback and that this was a cost and time effective way of providing a sizeable and available resource for teaching. A later study investigated the use of senior students to assess junior students during an OSCE examination. (Harris et al, 1990) In this study, students alternated as patients and examiners,
with no concurrent use of faculty examining the same candidates. The authors reported qualitative insights and determined the senior students were regarded as credible by the juniors, provided quick feedback to candidates, were readily available and trainable in one hour sessions, were reliable over the course of the day, and viewed their participation as an opportunity to develop teaching skills and enhance their own learning. A subsequent report of this project reported the motivation and benefits to the senior students (Davis Feickert et al, 1992) Students indicated they were motivated to participate by the honorarium and their interest in medical education, and viewed the experience as fun. Another group of investigators compared peer role-playing by medical students to SPs in evaluating the teaching of smoking cessation techniques. (Papadakis et al, 1997) Although candidates valued the SP experience more highly than role-playing, there was no difference between groups in performance in the subsequent clinical skills testing exercise. The authors concluded this method is less expensive but equally effective to using SPs to provide students with performance skills. This work contrasts with others who found students' role-playing was the only one of four major components of a smoking-cessation curriculum that was not rated by medical students and faculty as useful. (Ursatine et al, 1996) Finally, researchers who used senior medical students as SPs in a clinical practice exam for junior students found the seniors demonstrated superior scores to their inexperienced peers when their own communication skills were tested in a similar manner. (Sasson et al, 1999)

One author has reviewed the ethical acceptability of using medical students as surrogate patients, suggesting that ethical strengths and weaknesses depend upon whether students are involved for the sake of teaching or for the sake of research. The promotion of
free and informed involvement of students as surrogate patients was recommended. (Braunack-Mayer, 2001)

In the arena of peer role-playing and evaluations, senior medical students were accepting of peer evaluations (Helfer, 1972) and peer evaluations of professionalism correlated well with faculty evaluations of the same dimension. (Thomas et al, 1999) Medical students have demonstrated comfort in using peers vs SPs' as subjects for physical examinations (Chang et al 2000). There is also some evidence supporting training of students as standardized 'candidates' for quality control. (Ellen et al, 1994; Pangaro, 1997)

2.12 Impact of Participation on SPs

Few investigators have examined the impact of participating on the standardized patient. One reported that simulated patients get so involved in the role they have begun to imagine they really had symptoms (Davis, 1989) and another cautioned that SPs may not be able to separate their personal history or experience from the case they are simulating. (Barrows, 1971) Eight of 9 actors who regularly simulated patients disorders and 8 of 9 actors who simulated a disorder for the first time, suffered some emotional sequelae and the necessity of repeating them frequently was the most distressing aspect of the simulations. (Naftulin and Andrew, 1975) Other authors confirm this, suggesting the highly emotional nature of the OSCE roles can have residual effects on the SPs. (Hodges et al, 1997). In a pilot study of SPs portraying psychologically and emotionally complex standardized patient roles, 69% described residual psycho-physiological effects, while all reported some effect of portraying emotional roles, sometimes lasting several days. (McNaughton et al, 1999)
Residual effects may have been increased or mitigated by variables such as repetition, total time and acting style. Yet all were strongly motivated to continue to simulate psychiatric roles because of the social value of the work to society, and to the program, individual learning, benefit to acting skills, socializing and remuneration. Another study of the health care perceptions of standardized patients showed these worsened after 12 months of OSCE participation, but it was not apparent if the effects were attributable to better education for the consumer or variables in the OSCE. (Rubin & Philip, 1999)

Other researchers who examined the effect of simulating on actors as patients discovered they developed a more balanced perspective regarding health care professionals, improvement in their communication skills, and gains in tolerance or open-mindedness. (Woodward et al, 1995) Patients as SPs benefited from teaching, demonstrating higher quality of life and more satisfying relationships with their physicians. (Gecht, 2000) Teenage actors gained first-hand knowledge on health issues, while learning that physicians have doubts and debates, and that postgraduate continuous learning is part of their professional lives. (Hardoff & Schonmann, 2001)

Several studies have considered benefits to real patients from involvement in teaching. The major element seems to be the satisfaction of helping in the education and training of doctors and feeling better informed. (O'Malley et al, 1997; Thomas et al, 1999; Lynne et al, 1998; Spencer et al, 2000)

The literature describes benefits for medical students who teach other medical students. Students who peer-taught anatomy felt they would be better able to communicate with patients and felt more like students of medicine, while faculty observed students encouraged and supported each other. (Hendelman et al, 1986) Medical student tutors have
reported a benefit in preparation for medical licensure examinations (Walker-Bartvick et al, 1984) and the value in group leadership and being a teacher. (Resnick et al, 1976)

2.13 Summary

This literature review has outlined the development of performance-based assessment as an educational measurement system for clinical competency. It described the emergence of the objective structured clinical examination as the preferred form of PBA used in clinical competency assessment in the health professions arena and highlighted its components. It compared the typical assessment tools of binary checklists and global rating scales.

It then described the use of standardized patients in terms of the value to students, to teaching, to assessment and to program development. It examined fidelity issues of realism, reliability and validity in assessment; test security; and reviewed the literature comparing faculty vs SP raters. This was followed by a summary of the relevant literature on the use of students as standardized patients, and finally, the impact of participation on SPs. This literature exploration has thus presented the context for this investigation into the impact of junior students as standardized patients and raters in a senior candidate OSCE.

It is clear, in examining the literature, that the assessment of clinical competency has progressed rapidly in the past two decades, fueled by dramatic investigations surrounding the candidates, patients, raters, teaching methodology, assessment tools and format, and psychometric and qualitative examination of the impact of the evaluation. The results of a study of the ramifications of using first year pharmacy students as standardized patients on the psychometric properties of a senior OSCE can provide new insights into the continued
evolution of this arena and be used to plan further research and to formulate strategic educational plans and programs for the future.
Chapter 3: Research Design and Methodology

3.1 Description of the OSCE Used in the Study

This study examined the University of Toronto Faculty of Pharmacy 3rd year self-medication OSCE administered in March 2000. The OSCE is designed to measure the competencies acquired by candidates during completion of two sequential self-medication courses taught during second and third year. The cumulative exit OSCE occurs in March, 10 weeks after completion of the second course. The examination was administered over two successive days and tested 108 candidates. Each day, the candidate completed two ten-minute interviews followed by a 5-minute post encounter assessment period. There was an interval between stations of 45 to 60 minutes. To facilitate the process, five parallel circuits were administered in three consecutive sessions on each day. The first day and second day interviews were scheduled for each candidate at the same times on each day, so that a 24-hour time lapse occurred for each. Table 2a and 2b illustrates the OSCE circuits for each candidate for each day.

Following the circuit illustrated in Tables 2a, candidate number 1 can be used as an example of the test sequencing for each candidate outlined. This candidate had his first interview, case one, in room A, from 8:30 until 8:45. He was then escorted by an invigilator and sequestered in a holding room for forty-five minutes. At this time, the invigilator escorted him to room F, where he had his second interview, case two, from 9:30 until 9:45. At this time, he was escorted by an invigilator out of the building through a back entrance.
and returned the following day. Table 2b illustrates that the same timed sequence was repeated for this candidate on the second day, for cases seven and eight.

On day one, a standardized actor (SP) from the University of Toronto Standardized Patient Program played the patient. On day two, a first year junior pharmacy student (JR) from the Faculty of Pharmacy played the patient. The performance of candidates at each station was observed and rated by a practicing pharmacist (TA) who is a teaching assistant or professor at the Faculty. Table 3 outlines the patients and raters for the OSCE.

3.2 Quality Control of Consistency in Role Delivery

For each examination day, all patients and TA raters who participated in the same case watched a preliminary enactment of the role to establish consistency in role-delivery. The course instructor rotated through each set of five rooms assigned per case to observe and comment on consistency in role delivery and consistency in use of the assessment tool.

3.3 Description of the Data Collection

During the encounter, the TA observer completed a case-specific five item task-specific checklist, which related to problem identification and treatment. Following the encounter, the candidate left the room. The TA, the patient and the candidate, completed a global rating scale, which was the same for all stations. Patient and TA raters were carefully cautioned to resist the inclination to compare impressions. Their scores were included in the data analysis.
Table 2a:
The OSCE Circuits: Day One

<table>
<thead>
<tr>
<th>Time</th>
<th>Activities for Day One</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:45-8:15</td>
<td>Briefing</td>
</tr>
<tr>
<td>8:15-8:30</td>
<td>Session One</td>
</tr>
<tr>
<td></td>
<td>Practice run of case 1 or 2 with 1 simulated patient and 1 teaching assistant</td>
</tr>
<tr>
<td></td>
<td>Case 1</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>8:30-8:45</td>
<td>1</td>
</tr>
<tr>
<td>8:45-9:00</td>
<td>6</td>
</tr>
<tr>
<td>9:00-9:15</td>
<td>11</td>
</tr>
<tr>
<td>9:15-9:30</td>
<td>16</td>
</tr>
<tr>
<td>Switch Cases</td>
<td></td>
</tr>
<tr>
<td>9:30-9:45</td>
<td>21</td>
</tr>
<tr>
<td>9:45-10:00</td>
<td>26</td>
</tr>
<tr>
<td>10:00-10:15</td>
<td>31</td>
</tr>
<tr>
<td>10:15-10:30</td>
<td>36</td>
</tr>
<tr>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>10:45-11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Session Two</td>
</tr>
<tr>
<td></td>
<td>Practice run of case 3 or 4 with 1 simulated patient and 1 teaching assistant</td>
</tr>
<tr>
<td></td>
<td>Case 3</td>
</tr>
<tr>
<td>11:00-11:15</td>
<td>41</td>
</tr>
<tr>
<td>11:15-11:30</td>
<td>46</td>
</tr>
<tr>
<td>11:30-11:45</td>
<td>51</td>
</tr>
<tr>
<td>11:45-12:00</td>
<td>56</td>
</tr>
<tr>
<td>Switch Cases</td>
<td></td>
</tr>
<tr>
<td>12:00-12:15</td>
<td>61</td>
</tr>
<tr>
<td>12:15-12:30</td>
<td>66</td>
</tr>
<tr>
<td>12:30-12:45</td>
<td>71</td>
</tr>
<tr>
<td>12:45-1:00</td>
<td>76</td>
</tr>
<tr>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>1:45-2:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Session Three</td>
</tr>
<tr>
<td></td>
<td>Practice run of case 5 or 6 with 1 simulated patient and 1 teaching assistant</td>
</tr>
<tr>
<td></td>
<td>Case 5</td>
</tr>
<tr>
<td>2:00-2:15</td>
<td>81</td>
</tr>
<tr>
<td>2:15-2:30</td>
<td>84</td>
</tr>
<tr>
<td>2:30-2:45</td>
<td>87</td>
</tr>
<tr>
<td>2:45-3:00</td>
<td>90</td>
</tr>
<tr>
<td>3:00-3:15</td>
<td>93</td>
</tr>
<tr>
<td>Switch Cases</td>
<td></td>
</tr>
<tr>
<td>3:15-3:30</td>
<td>96</td>
</tr>
<tr>
<td>3:30-3:45</td>
<td>99</td>
</tr>
<tr>
<td>3:45-4:00</td>
<td>102</td>
</tr>
<tr>
<td>4:00-4:15</td>
<td>105</td>
</tr>
<tr>
<td>4:15-4:30</td>
<td>108</td>
</tr>
<tr>
<td>4:30-4:50</td>
<td></td>
</tr>
<tr>
<td>De-Briefing</td>
<td></td>
</tr>
</tbody>
</table>
Table 2b:
The OSCE Circuits: Day Two

<table>
<thead>
<tr>
<th>Time</th>
<th>Activities for Day Two</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:45-8:15</td>
<td>Briefing</td>
</tr>
<tr>
<td>8:15-8:30</td>
<td>Session One</td>
</tr>
<tr>
<td>8:30-8:45</td>
<td>Practice run of case 7 or 8 with 1 simulated patient and 1 teaching assistant</td>
</tr>
<tr>
<td>Room</td>
<td>Candidate Number</td>
</tr>
<tr>
<td>A</td>
<td>1  2  3  4  5  21  22  23  24  25</td>
</tr>
<tr>
<td>B</td>
<td>6  7  8  9  10  26  27  28  29  30</td>
</tr>
<tr>
<td>C</td>
<td>11 12 13 14 15  31  32  33  34  35</td>
</tr>
<tr>
<td>D</td>
<td>16 17 18 19 20  36  37  38  39  40</td>
</tr>
<tr>
<td>9:30-9:45</td>
<td>Switch Cases</td>
</tr>
<tr>
<td>10:00-10:30</td>
<td>Session Two</td>
</tr>
<tr>
<td>10:45-11:00</td>
<td>Practice run of case 9 or 10 with 1 simulated patient and 1 teaching assistant</td>
</tr>
<tr>
<td>Case 9</td>
<td>41 42 43 44 45 61 62 63 64 65</td>
</tr>
<tr>
<td>Case 10</td>
<td>46 47 48 49 50 66 67 68 69 70</td>
</tr>
<tr>
<td>11:00-11:15</td>
<td>Switch Cases</td>
</tr>
<tr>
<td>12:00-12:15</td>
<td>Lunch</td>
</tr>
<tr>
<td>1:45-2:00</td>
<td>Session Three</td>
</tr>
<tr>
<td>2:00-2:15</td>
<td>Practice run of case 11 or 12 with 1 simulated patient and 1 teaching assistant</td>
</tr>
<tr>
<td>Case 11</td>
<td>81 82 83 96 97 98</td>
</tr>
<tr>
<td>Case 12</td>
<td>84 85 86 99 100 101</td>
</tr>
<tr>
<td>3:15-3:30</td>
<td>Switch Cases</td>
</tr>
<tr>
<td>4:00-4:15</td>
<td>De-Briefing</td>
</tr>
<tr>
<td>4:30-5:00</td>
<td></td>
</tr>
</tbody>
</table>
Table 3:
Patients and raters for the OSCE

<table>
<thead>
<tr>
<th>Stations</th>
<th>Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Station 1</td>
<td>SP rater #1</td>
</tr>
<tr>
<td>Station 2</td>
<td>SP rater #2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stations</th>
<th>Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Station 1</td>
<td>JR rater #1</td>
</tr>
<tr>
<td>Station 2</td>
<td>JR rater #2</td>
</tr>
</tbody>
</table>

Candidate self-assessments were not included in the data analysis. They were carried out in order to give the candidate an opportunity to reflect upon and evaluate their performance after each interview. It was an opportunity to garner insights in preparation for the next test interview.
At the completion of the four station OSCE, each candidate completed a survey to collect data regarding their impressions of the OSCE. These data were not included in this data analysis but were collected for future research.

3.4 Method for Recording Incidents

Each interview was audio-taped by the TA observer. They noted on the schedule any untoward events, the candidate involved and the tape number and time for future review and scrutiny.

In their self-assessments, the candidate also had the opportunity to record any untoward events and their impressions of the incident for future evaluation.

3.5 Case Design

A total of twelve cases were written for the OSCE. Six were used each day in pairs over three administered sessions. Each sessional pairing was designed to be similar in terms of complexity of the content, and complexity of the patient. However to ensure content validity, the clinical scenarios presented as wide a range of situations as a self-medication advisor would be expected to handle. The sampling was twelve of thirty curriculum topics or 40%. Each case was an interaction between a simulated patient and the candidate, who assumed the role of a practicing pharmacist. The cases were designed to address the course objectives in terms of content, processing the delivery of pharmaceutical care, and interpersonal skills, in addition to assisting patients with special needs. Cases were structured
to test competency in four separate areas of professionalism which the candidate was expected to identify and solve. These included:

- a communication challenge
- a special need, such as a social, emotional or cognitive issue
- a key component in history-taking relating to medical conditions, family history, allergies or medications
- the identification, prevention or resolution one or two drug-related problems

These key points were summarized in a short, five point checklist of task-specific items that would be required of the candidates relating to problem identification and treatment. Integration of this knowledge with skills was required to pass the interview and was reflected in the scoring of the assessment tool in the domain that addressed overall impression (knowledge and skills).

The candidate was expected to dialogue with the patient effectively. The level of critical thinking required for each case required Bloom’s higher order skills, including analysis, synthesis and evaluation built upon recall, understanding, and application of course material. The scope of the material was cumulative over the two courses. The cases were reflective of realistic patients in pharmacy practice. This content validity was assured in that they were real patient situations encountered by the course instructor in her pharmacy practice as an advisor for self-medication. In that practice site, the instructor has no other tasks except dialoguing with patients in the over-the-counter (OTC) sections in a community pharmacy and typically will advise 120-130 patients during each shift. The cases were written by the case instructor and reviewed by another instructor for content and level of
difficulty. A template developed by the Doctor of Pharmacy program at the Faculty of Pharmacy, University of Toronto, was employed to develop the cases in a standardized format. Appendix Two is the template used.

3.6 Description of Assessment Tool: a Global Rating Scale

The assessment tool used to gather data for analysis was a global rating scale, which was the same for all stations and used by all raters. The tool was designed by a rhetorician in consultation with clinical educators from several health care disciplines to test processing skills which reflect competence. The scale consists of four domains, which assess empathy, coherence (organization/focus), verbal skills and non-verbal skills on a 5-point scale. A fifth domain is the overall impression of the integration of knowledge and skills in the performance, also on a 5-point scale. Formal evaluation has shown that this instrument has quite acceptable psychometric properties. (Hodges et al, 1997) The five point scale is aligned to differentiate levels of processing ranging from novice, beginner, competent, proficient and expert and rewards efficiency and mastery rather than thoroughness. Appendix Three is the global rating scale used for the study OSCE.

3.7 Description of the OSCE Practices

Candidates were provided with extensive information about the purpose and format of the testing in advance of the OSCE. This was to prevent candidate misconceptions about level and quality of performance expected by the test design, and to allow formative
feedback to enhance performance skills. At the commencement of the second year course, a training session introduced students to the global rating scale. The purpose, use and interpretation were outlined using videotapes of student interviews.

Three types of formative practices were experienced by candidates prior to the OSCE. First, after each hour case-based class on a self-medication topic in both second and third year, they observed or participated as a pharmacist in ten-minute practice interviews conducted in front of the entire class. The patient was a junior student and the case, an applied version of the same topic, was written by a class team. The global rating scale was scored by the junior, the pharmacist and the team who wrote the case. Feedback comments were elicited from each of the raters and a class discussion was generated. Secondly, a formal practice of ten-minute interviews was conducted as part of the 3rd year Pharmacy Practice laboratory, in teams of 6 students with a TA. The students role-played patients for each other in turn, and the cases were written by the course instructor for self-medication. The global rating scale was used by each observer which generated discussion and suggestions for improvement. Lastly, each candidate had a private ten-minute practice interview with a junior student prior to the OSCE. The cases were written by the self-medication course instructor. The junior student rated the candidate using the global rating scale and gave immediate oral feedback regarding their feelings as a patient during the dialogue; and the organization, focus, and communication skills of the candidate. Each junior repeated the role-portrayal with two or three candidates.
3.8 Description of the Study Participants

3.8.1 Description of the Candidates

The 108 candidates were the 3rd year undergraduate cohort of OT1, Faculty of Pharmacy, University of Toronto. They were students in the sequential self-medication courses that prepare students to acquire the knowledge, skills, attitudes and attributes of professional practice, reflective practice, clinical problem-solving, teamwork, empowerment and leadership required for competency as advisors to patients for over-the-counter drugs. They had successfully completed the first problem-based learning (PBL) self-medication course, PHM 220 in their 2nd year, by satisfying the requirements of peer-teaching, team learning, completion of case-based assignments and two written essay examinations which tested content knowledge on fifteen conditions. They had completed the second PBL self-medication course, PHM 320, in the fall of 1999, by satisfying the requirements of peer-teaching, team learning, completion of case-based assignments, design of a consumer website lesson on one topic, and two case-based written essay examinations which tested content knowledge on sixteen cases. These requirements constituted 75% of their grade for this course. Their score on the OSCE constituted the remaining 25% of their grade for the course.

3.8.2 Description of the Standardized Patients (SP)

The thirty standardized actors were selected from the University of Toronto Standardized Patient Program. A total of six patient roles were delivered by SPs. These actors received careful training for their roles by the program trainer. The roles were refined by role-playing with the course professor, who is the investigator for this study. The trainer
then trained each SP for consistency in delivery. The SPs were also trained for consistency in the use of the global rating scale scoring tool by the trainer and the course professor. Each SP repeated the assigned role for no more than eight ten-minute interviews on the examination day. More than one SP per case was trained lest fatigue developed after four or five encounters with roles that were specifically demanding.

3.8.3 Description of the Junior Students (JR)

The 140 first year students were the undergraduate cohort of OT3, Faculty of Pharmacy, University of Toronto. They were students in a pharmacy social science course, which examines the perspective of the patient. These students had also completed a full semester course in communication skills. Sixty students who scored highest on their entrance to pharmacy test of oral and written communications skills were selected for participation in the exam. A total of six patient roles were delivered by JRs. They received their roles, but no description of content checklist items, ten days in advance. Juniors were then trained by the self-medication instructor both for consistency in using the global rating scale, using videotapes of student interviews, and in standardizing their roles. Each junior repeated the assigned role for no more than four ten-minute interviews on the examination day, in order to utilize all sixty students, and to protect against fatigue in these inexperienced role-players.

The remainder of the junior class was sorted into sections for participation in either the class practices (one per student) or the private individual interviews (two to three per student). Participation in the OSCE or the practices was part of their curriculum cross-course requirements. After each experience, juniors completed a survey to collect data regarding
their impressions. These results were not included in the data analysis but were collected for future research.

3.8.4 Description of the Pharmacist Examiners (TAs)

The examiners were practicing pharmacists who were either teaching assistants (TAs) or professors at the Faculty of Pharmacy, University of Toronto. They were trained by the self-medication course instructor for consistency in using the global rating scale, in group teaching sessions using videotapes of student interviews. These raters also received a detailed summary of the content covered in each test case, acceptable options for solving each case, and a five point checklist of task-specific items that would be required of the candidates relating to problem identification and treatment. Prior to the examinations, these TA raters and the course instructor reviewed the content of the cases together and agreed on which key items would be required for a pass.

3.9 Confidentiality Issues

Prior to the examination, each candidate and each first year student were required to read the code of conduct governing test security for examinations at the University. They were then required to sign a confidentiality agreement, which prohibited any discussion of the nature of the patient roles or stations until the examination concluded. The literature review of test security using SPs concluded that there is little problem with test security, especially when the SP application is measuring process rather than the ability to recall
content and produce answers, and that security should be monitored, but the issues should not restrict use. (Stillman, 1993) It was hoped that operationally, an element of uncertainty among the candidates was present as to the identity of the cases, and the use of 12 different cases and stringent efforts to enforce security would discourage any dissemination of confidential checklist items by first year students.

During the examination, candidates were sequestered in separate rooms with invigilators for each parallel stream to ensure there was no verbal contact. Incoming candidates were sequestered before outgoing candidates left the building. Each candidate and first year student were required to disclose in advance, any close relationship with a student in the other year, and the pairings were scheduled to avoid combining first year students and candidates who may have known one another.

3.10 Ethical Review

An ethical review was conducted according to the procedures and requirements of the Department of Adult Education, Teaching and Learning at The Ontario Institute for Studies in Education of the University of Toronto.

3.10.1 Ethical and Methodological Concerns

The examination that used for the purposes of the study occurred in March, 2000.

For all candidates who took the examination, the OSCE was part of their course requirements, and an examination of data a year after they had exited from the course was
part of the quality control for the course. This analysis did not affect their marks in any way, and all individual names and student numbers were removed from the data set.

For all first year students who participated, the OSCE was part of their curriculum cross-course requirements, and an examination of data a year after they had exited from the course was part of the quality control for the course. This analysis did not affect their marks in any way, and no individual names or student numbers were recorded for the data set.

3.11 Data Analysis

"Analysis of any kind involves a way of thinking. It refers to the systematic examination of something to determine its parts, the relationship among parts and their relationship to the whole. Analysis is a search for patterns." (Spradley, 1979, p 85)

The impact of using first year students as patients will be examined under two considerations for data analysis: impact on the faculty raters, and impact on patient rater.

(Refer to Appendix One for Summary)

3.11.1 Impact on Faculty Raters:

To examine the effect on reliability, an Olkins Z test will be used to compare inter-station correlations (station 1 vs station 2) generated by TAs raters when using SPs as patients and when using JRs as patients. To see if there is a systematic influence on candidate scores, a paired T-test will compare the mean of scores generated by TAs raters when using SPs as patients to the mean when using JRs as patients.
To examine concurrent validity, Pearson correlations will be used to compare average scores generated by TAs when SPs were patients and when JRs were patients to other measures of student performance. These include:

- self-medication course 3rd year written examination scores (case-based essay type)
- oral exam scores in a 3rd year therapeutics course (case-based, question/answer)
- GPA from current year (2000)
- GPA from previous year (1999)
- 3rd year traditional science courses
- clinical scores on fourth year exit course (2001) – predictive validity

3.11.2 Impact on Patient Raters:

These analytic tests will be repeated to assess the impact on patient raters using SPs vs JRs.

To examine the effect on reliability, an Olkins Z test will be used to compare inter-station correlations (station 1 vs station 2) generated by SP raters to inter-station correlations (station 1 vs station 2) generated by JR raters. To see if there is a systematic influence on candidate scores, a paired T-test will compare the mean of scores generated by SP raters to the mean of scores generated by JR raters.

To examine concurrent validity, Pearson correlations will be used to compare average scores generated by SPs and average scores generated by JRs to other measures of student performance. These include:
- self-medication course 3rd year written examination scores (case-based essay type)
- oral exam scores in a 3rd year therapeutics course (case-based, question/answer)
- GPA from current year (2000)
- GPA from previous year (1999)
- 3rd year traditional science courses
- clinical scores on fourth year exit course (2001) – predictive validity

In addition, to assess concurrent validity, the correlations between SP vs TA scores and JR vs TA scores will be assessed for station 1 and 2.
Chapter 4: Results

For the 108 candidates examined, the overall examination score was 73.1% with a standard deviation of 6.7 and a range from 53.3% to 86.5%. Four students failed (3.7%) and 15 (14%) received honour-status marks. There were no untoward events reported, and there was no suggestion that confidentiality or station security was a problem.

4.1 Impact on Faculty Raters:

4.1.1 Inter-station Correlations

Data were divided into the three sessions in the OSCE circuits and compared for each session and for combined sessions, to see if there was an effect on inter-station correlations of TA scores when using JRss vs SPs as patients. These are reported in Table 4. The correlations in TA scores between station 1 and station 2 when the SP role-played the patient and when the JR played the patient, were not statistically different. An Olkin’s Z test was therefore not performed. This lack of correlation may not be a function of rater consistency or test equivalency (not all candidates see the same cases) but rather a function of the students’ abilities.
Table 4: Reliability

Correlations of TA Scores between Station 1 and Station 2 for the Pairs of Stations when SPs and JRs were used as patients: *

<table>
<thead>
<tr>
<th>Patient</th>
<th>Pair 1</th>
<th>Pair 2</th>
<th>Pair 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPs</td>
<td>.293</td>
<td>.090</td>
<td>.076</td>
</tr>
<tr>
<td>JRs</td>
<td>-.015</td>
<td>-.050</td>
<td>.188</td>
</tr>
</tbody>
</table>

*no values were statistically significant (all p values >.05)

4.1.2 Influence on Candidate Scores

A paired t-test comparing means was performed to examine if the use of JRs vs SPs as patients had a systematic influence on candidates scores. Data analysis showed there were no significant differences in the TA scores ($t_{104} = 1.84, p = .07$) when the SPs played the patient, ($\bar{x} = 4.26 \pm 1.07$) vs when JRs played the patient ($\bar{x} = 4.49 \pm 0.96$), with the difference in means amounting to less than 0.23 standard deviation size. This may be evidence of the two tests' equivalence since the mean performance scores were similar.

4.1.3 Concurrent Validity

To test concurrent validity, Pearson correlations were used to compare average scores generated by TAs when SPs vs JRs were patients, to other measures of student performance.

The first row in Table 5 tabulates these results when SPs were used patients. The Pearson correlations were low but statistically significant between the TA OSCE scores and the following: the oral exam in therapeutics ($r = 0.244, p = .011$); the GPA from the current year ($r =
0.299, \( p = .002 \)); and with the GPA from the previous year \((r = 0.393, p = .0001)\). Three other low but statistically significant Pearson correlations were found between TA OSCE scores and the self-medication course (PHM 320) 2 written exams \((r = 0.440, p = .0001)\); the averaged traditional science courses \((r = .313, p = .001)\); and the next year’s exit clinical course, which were one hospital and one community pharmacy rotation of two months each, \((r = .231, p = .017)\). These were then adjusted in consideration of the reliability of these three compared measure, calculated as follows.

It is worth noting that the scores being compared or predicted are themselves somewhat unreliable (error prone). It is possible to calculate what the correlation with these measure would be if they were perfectly reliable using the correction for attenuation formula. This formula is based on the recognition that the reliability \((r_{xx})\) of the measure acts as a theoretical ceiling of its ability to correlate with another measure and adjusts the observed correlation to account for this attenuation. To make this adjustment, the observed correlation is divided by the square root of the reliability of the measure. For measures where a reliability \((r_{xx})\) coefficient or Cronbach’s alpha could be calculated, row 3 of Table 5 presents the reliability \((r_{xx})\) and row 5 presents the disattenuated or adjusted correlation.
Table 5: Concurrent Validity

TA Scores Using SPs as Patients Compared to other Measures:

<table>
<thead>
<tr>
<th></th>
<th>Two Written OTC exams</th>
<th>Oral Exam Therapeutics</th>
<th>Traditional Science Courses</th>
<th>GPA 2000</th>
<th>GPA 1999</th>
<th>Clinical Exit Marks 2001</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pearson</strong> correlation</td>
<td>.440**</td>
<td>.244*</td>
<td>.313**</td>
<td>.299**</td>
<td>.393**</td>
<td>.231*</td>
</tr>
<tr>
<td><strong>Sig (2-tailed)</strong></td>
<td>.0001</td>
<td>.011</td>
<td>.001</td>
<td>.002</td>
<td>.0001</td>
<td>.017</td>
</tr>
<tr>
<td><strong>Reliability of 2nd measure</strong></td>
<td>.77</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
<td>.3998</td>
</tr>
<tr>
<td><strong>Square root of 2nd measure</strong></td>
<td>.877</td>
<td>.938</td>
<td></td>
<td></td>
<td></td>
<td>.632</td>
</tr>
<tr>
<td><strong>Adjusted Correlation (row1/row 4)</strong></td>
<td>.50</td>
<td>.333</td>
<td></td>
<td></td>
<td></td>
<td>.365</td>
</tr>
</tbody>
</table>

The first row of Table 6 tabulates the results when JRs were used patients. A low but statistically significant Pearson correlation ($r = 0.198$, $p = 0.041$) was found with the oral exam in therapeutics. Two other low but statistically significant correlations were found with the self-medication course (PHM 320) 2 written exams ($r = 0.191$, $p = 0.049$) and the next year’s exit
clinical course \((r = 0.299, p = .002)\) For measures where the reliability of the compared measure could be calculated, row 3 of Table 6 presents the reliability and row 5 of Table 6 presents the disattenuated or adjusted correlation. No statistically significant correlations were found between TA scores using JRrs as patients and the traditional science courses average, \((r = 0.13, \text{not significant})\) the GPA from the current year, \((r = 0.143, \text{not significant})\) or the GPA from the previous year. \((r = 0.149, \text{not significant})\)

4.2 Impact on Patient Raters:

4.2.1 Inter-station Correlations

Data were divided into the three sessions in the OSCE circuits and compared for each session and for combined sessions, to see if there was an effect on inter-station correlations of SP scores JRrs scores. For the SP raters, there was a small but significant correlation between stations in session one \((r = 0.364, p = .02)\) and session 2 \((r = 0.365, p = .02)\) but correlations were not significant for session 3 nor for the combined sessions. Correlations for JR scores between stations were not statistically significant. These are reported in Table 7. An Olkin’s Z test was therefore not performed. Once again, this lack of correlation may not be a function of rater consistency or test difficulty but rather a function of the students’ abilities.
Table 6: Concurrent Validity

TA Scores Using JRs as Patients Compared to other Measures:

<table>
<thead>
<tr>
<th></th>
<th>Two Written OTC Exams</th>
<th>Oral Exam Therapeutics</th>
<th>Traditional Science Courses</th>
<th>GPA 2000</th>
<th>GPA 1999</th>
<th>Clinical Exit Marks 2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation</td>
<td>.191*</td>
<td>.198*</td>
<td>.045</td>
<td>.046</td>
<td>.097</td>
<td>.299**</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
<td>.049</td>
<td>.041</td>
<td>.648</td>
<td>.637</td>
<td>.322</td>
<td>.002</td>
</tr>
<tr>
<td>Reliability of 2nd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.3998</td>
</tr>
<tr>
<td>measure</td>
<td>.77</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Square root of 2nd</td>
<td>.877</td>
<td>.938</td>
<td></td>
<td></td>
<td></td>
<td>.632</td>
</tr>
<tr>
<td>measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted Correlation</td>
<td>.218</td>
<td></td>
<td></td>
<td></td>
<td>.473</td>
<td></td>
</tr>
<tr>
<td>(row1/row 4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7: Reliability of SP and JR Raters

Correlations of Scores between Station 1 and Station 2 for the Pairs of Stations when SPs were raters and when JRs were raters: *

<table>
<thead>
<tr>
<th>Rater</th>
<th>Pair 1</th>
<th>Pair 2</th>
<th>Pair 3</th>
<th>Combined Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPs</td>
<td>.364* (p = .05)</td>
<td>.365* (p = .05)</td>
<td>.019</td>
<td>.136</td>
</tr>
<tr>
<td>JRs</td>
<td>.157</td>
<td>-.155</td>
<td>.249</td>
<td>.127</td>
</tr>
</tbody>
</table>

*no values were statistically significant for the combined sessions (all p values > .05)

4.2.2 Influence on Candidate Scores

A paired t-test comparing means was done to examine if the use of SPs vs JRs as patients had a systematic influence on candidates scores. Data analysis showed there were no significant differences (t₁₀₄ = 0.27, p = .79) in the mean SP scores (\( \bar{x} = 4.93 \pm .88 \)) and the mean JR scores (\( \bar{x} = 4.90 \pm .83 \)), with the difference in means equaling approximately 0.05 standard deviations. This indicated that both kinds of patient raters marked similarly.
4.2.3 Concurrent Validity

To test concurrent validity, Pearson correlations were used to compare average scores generated by SPs and by JRs, to other measures of student performance.

The first row of Table 8 tabulates these results when SPs rated the candidates. The Pearson correlations were low but statistically significant between the SP OSCE scores and the following: the oral exam in therapeutics ($r = 0.231, p = .017$); the GPA from the current year ($r = 0.271, p = .005$); and the GPA from the previous year ($r = 0.263, p = .006$). Two other low but statistically significant correlations were found between SP average scores and the self-medication course (PHM 320) written exams ($r = 0.257, p = .008$) and the averaged traditional science courses ($r = .262, p = .005$). These were then adjusted in consideration of the reliability of the compared measure. Row 3 of Table 8 presents the reliability and row 5 of Table 8 presents the disattenuated or adjusted correlation. No statistically significant correlations were found between SP scores and the clinical exit course scores for the following year. ($r = 0.136$, not significant)

The first row of Table 9 tabulates these results when JRs rated the candidates. The Pearson correlations were low but statistically significant between the JR OSCE scores and the following: the oral exam in therapeutics ($r = 0.285, p = .003$); and the next year's exit clinical course ($r = 0.240, p = .03$). Row 3 of Table 9 presents the reliability and row 5 of Table 9 presents the disattenuated or adjusted correlation. No other statistically significant correlations were found between JR scores and other measures.
Table 8: Concurrent Validity

SP Scores Compared to other Measures:

<table>
<thead>
<tr>
<th></th>
<th>Two Written OTC Exams</th>
<th>Oral Exam Therapeutics</th>
<th>Traditional Science Courses</th>
<th>GPA 2000</th>
<th>GPA 1999</th>
<th>Clinical Exit Marks 2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation</td>
<td>.257**</td>
<td>.262**</td>
<td>.271**</td>
<td>.263**</td>
<td>.136</td>
<td></td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
<td>.008</td>
<td>.017</td>
<td>.007</td>
<td>.005</td>
<td>.006</td>
<td>.166</td>
</tr>
<tr>
<td>Reliability of 2\textsuperscript{nd} measure</td>
<td>.77</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
<td>.3998</td>
</tr>
<tr>
<td>Square root of 2\textsuperscript{nd} measure</td>
<td>.877</td>
<td>.938</td>
<td></td>
<td></td>
<td></td>
<td>.632</td>
</tr>
<tr>
<td>Adjusted Correlation (row1/row 4)</td>
<td>.293</td>
<td>.279</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 9: Concurrent Validity

JR Scores Compared to other Measures:

<table>
<thead>
<tr>
<th></th>
<th>Two written OTC exams</th>
<th>Oral Exam Therapeutics</th>
<th>Traditional Science Courses</th>
<th>GPA 2000</th>
<th>GPA 1999</th>
<th>Clinical Exit Marks 2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation</td>
<td>.155</td>
<td>.285**</td>
<td>.130</td>
<td>.143</td>
<td>.149</td>
<td>.240**</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
<td>.110</td>
<td>.003</td>
<td>.181</td>
<td>.142</td>
<td>.125</td>
<td>.013</td>
</tr>
<tr>
<td>Reliability of 2nd measure</td>
<td>.77</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
<td>.3998</td>
</tr>
<tr>
<td>Square root of 2nd measure</td>
<td>.877</td>
<td>.938</td>
<td></td>
<td></td>
<td></td>
<td>.632</td>
</tr>
<tr>
<td>Adjusted Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.380</td>
</tr>
</tbody>
</table>

4.3 Concurrent Validity:

As an additional measure to assess concurrent validity, the correlations between SP vs TA scores and JR vs TA scores were assessed for station 1 and 2. These are tabulated in Table 10. The Pearson correlations for the SP vs TA scores were 0.378 (p = .0001) for station 1 and
0.579 (p = .0001) for station 2. The Pearson correlations for the JR vs TA were 0.584 (p = .0001)
for station 1 and 0.619 (p = 0.0001) for station 2. The correlations for the averaged scores for TA
vs SP raters was 0.565 (p = .0001) and f0TA vs JR raters was 0.646 ( p = .0001). These strong
Pearson correlations verified that the TA and JR scores were statistically similar for each station
as were the TA and SP scores. Therefore, SP and JR raters produced correlations that were both
reasonable and the same compared to TA raters.

However, it is worth noting that the means of both SP (\(\bar{x} = 4.93 \pm .88\)) and JR (\(\bar{x} = 4.90 \pm
.83\)) raters were higher than the means of TA raters, (SP patients, \(\bar{x} = 4.26 \pm 1.07\); JR patients, \(\bar{x} =
4.49 \pm 0.96\)) and to the same level. Thus both types of patient raters marked equally more
generously as compared to TA raters.
Table 10: Concurrent Validity

Correlations of TA/SP, TA/JR and Averaged Scores for Stations 1 and 2

<table>
<thead>
<tr>
<th></th>
<th>TA/SP Scores</th>
<th>TA/JR Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Station 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>.378</td>
<td>.584</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
<td>.0001</td>
<td>.0001</td>
</tr>
<tr>
<td>Station 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>.579</td>
<td>.619</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
<td>.000</td>
<td>.0001</td>
</tr>
<tr>
<td>Averaged Scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>.565</td>
<td>.646</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
<td>.0001</td>
<td>.0001</td>
</tr>
</tbody>
</table>
Chapter 6: Discussion and Conclusions

6.1 Introduction

This final chapter provides summaries of the data analysis in terms of conclusions and implications from the research. Recommendations for future study endeavours and final comments are provided.

6.2 Purpose of the Study

The purpose of this study was to use the third year self-medication OSCE at the University of Toronto, Faculty of Pharmacy to study the impact of using first year junior students as standardized patients, by examining the psychometric properties of reliability and validity of using first year student, SP and faculty raters to evaluate performance.

6.3 Summary of the Study

This research investigated an existing database of scores generated in the March 2000 OSCE for PHM 320 at the Faculty of Pharmacy. Statistical tests were performed on raw scores, with candidate identifiers removed. The quantitative analysis focused on addressing questions relating to impact on faculty raters, impact on patient raters and impact on
performance in terms of the reliability and validity of the examination scoring when using juniors vs standardized patients as both actors and raters.

6.4 Research Questions and Answers

Is there an impact on Faculty Raters (TA) Scores of Candidate Performance when using junior students vs. SPs?

1. Reliability

- Is there a difference in TA scores using junior students vs SPs when correlations between stations are compared?

This analysis did not produce results from which any firm conclusions may be drawn. No significant inter-station correlations were seen for TA scores when either the SPs or the JRs were used as patients, and thus there was way to assess the effect of patient type or inter-station reliability. While the assumption could be made that this reflects a lack of consistency in the rating by the TAs, other variables could have been responsible. Low inter-station reliability is expected when using only two stations, and indeed the two cases per pairing were designed to sample disparate topics. Research has shown the reproducibility for domain-referenced score interpretation is lower due to effects of differences in the difficulty of stations. (van der Vleuten & Swanson, 1990) In addition to differences in the complexity of the content of the cases, differences in the complexity of the patient roles, variability in the role-portrayal by the actors across five parallel circuits, and a difference in the candidates' individual abilities as reflected in their performance all may have had an influence on the data produced. This is supported in well-articulated literature. Reliability analyses consistently indicate that the major source of
measurement error is variation in examinee performance from station to station (content specificity) which can be overcome with larger numbers of stations. Disagreements among raters observing examinee performance and differences between SPs playing the same patient role appear to have less effect on the precision of scores, as long as examinees are randomly assigned to raters and SPs (Swanson & Norcini, 1989; van der Vleuten & Swanson, 1990).

- **Is there is a systematic influence on candidate scores when the means of TA scores are compared for juniors vs SPs?**

  Data analysis showed there were no significant differences in the means of the TA scores between stations when the SPs played the patients vs when JRIs played the patient. This may be evidence that there is no significant impact on candidate scores when JRIs are used as patients.

2. **Concurrent validity:**

- **Do TA scores using juniors as patients show correlations with other measures of performance?**

  These data showed that when TAs rated the candidates, and SPs were used as patients, the OSCE showed significant, although low, correlations with other measures of performance, including the two written exam in the same course, a question and answer format oral exam in therapeutics, averaged traditional science course, and GPAs from current and previous years, and the averaged results of two clinical experience rotations held the following year as an exit requirement. In contrast, when JRIs were used as patients, there were low but significant correlations with fewer other performance measures, notably the two written exam in the same
course, and the oral therapeutics exam. Interestingly, there was a stronger correlation with the following clinical experiential course when juniors were patients than occurred with SPs as patients. Conclusions might be drawn that while using either SPs or JRs as patients produce valid results with concurrent and prior performance measures when TAs are raters, these effects are more generalizeable using SPs. Conversely, when TAs rate candidates’ performance using JRs as patients, results seem to show better predictive validity.

*Is there an impact on Patient Rater Scores of Candidate Performance when using junior students vs. SPs*

1. **Reliability**

   • *Is there a difference in patient scores, juniors vs SPs, when correlations between stations are compared?*

   Once more, these data did not produce results from which conclusions may be drawn. Although some small, yet significant correlations were seen between stations when SPs scored in sessions 1 and 2, these results were not repeated in the third session, or for the combined analysis. Thus, the correlations were highly unstable and not susceptible to comparison when either the SPs or the JRs were used as patients. Again, while this may reflect lack of consistency in the rating by the SPs or the JRs, other variables previously listed may play a role, including small number of stations, case variability, role-portrayal or candidates’ abilities.

   • *Is there a systematic influence on student scores when the means of juniors scores vs SPs scores are compared?*
There were no significant differences in the means of the SP scores and the JR scores between stations. However, both SPs and JR rated the candidates more generously than did the TAs. This suggests that while there is no significant impact in using juniors as opposed to SPs on candidate scores, there is an impact on candidate scores generated by simulated patients vs TAs. This reinforces other evidence and may be the reason that physician raters are believed to be more credible in a high-stakes exam (promotion or graduation) (Stillman, 1993). This may reflect the lack of content awareness of patients in contrast to trained pharmacist professionals. It may also be due to the increased emphasis on process skills of empathy, coherence and communication as seen from the patient’s perspective. From the patient’s subjective lens, these skills may be what convey a sense of trustworthiness and competence.

2. Concurrent validity:

- Do the the SPs scores and juniors scores show correlations with other measures of performance?

These data showed that when SPs rated the candidates, there were significant, although low, correlations with other measures of performance, including the two written exam in the same course, the oral exam in therapeutics, the GPAs from the current and previous year, and the averaged traditional science courses. However, no statistically significant correlations were found between SP scores and the clinical exit course scores for the following year. In contrast, when JR rated the candidates, there were low but significant correlations with one concurrent measure, the oral exam in therapeutics and the next year’s exit clinical course. While this evidence is somewhat weak, conclusions might be drawn that SPs raters produced valid results
with concurrent and prior performance measures, but were not predictive of candidates’ performance. In contrast when JRs rated candidates, these scores were valid compared to a single concurrent measure yet showed predictive validity for a clinical exit course. It is interesting to note that these two other comparators, the oral therapeutic exam and the clinical experiential course, are highly focused on measurement of processing skills of communication and organization while the other tests focus more on content. In fact, the comparison to the clinical experiential course relates the behaviours captured in the OSCE interaction to actual patient outcomes based on expert judgments of student performance: it is a good measurement of predictive validity and as such is a standard setting tool. One might argue that juniors appear to be valid raters of process skills compared to similar testing indices. This is aligned with literature reports that within domain-referenced and mastery-testing frameworks, the equivalence of SP-based and traditional tests disappears, because the absolute level of performance is of interest (Newble & Swanson, 1988) and that SP tests measure important aspects of clinical competence not tapped by traditional measures. (van der Vleuten and Swanson, 1990)

Is there a correlation in the SP vs TA scores and the junior vs TA scores between stations 1 and 2?

Finally, strong Pearson correlations verified that the TA and JR scores were statistically similar for each station as were the TA and SP scores, which indicate that TAs, SPs and JRs are similarly consistent in differentiating students of contrasting abilities and that using either SPs or JRs as raters is valid in comparison to professional TAs, when scoring with a global rating tool. This supports the literature which documents adequate inter-rater agreement can be achieved through use of either faculty or SPs as raters. (van der Vleuten and Swanson, 1990)
6.5 Implications and Conclusions

"...the goal is not just to optimize reliability, but rather to create an examination reflective of the programme or curricular objectives and then to ensure that its measurement methods and qualities are appropriate." Dauphinee & Blackmore, 2001

The concordance of curricular objectives and assessment should be paramount. For this study, defining what was to be tested was a critical phase, including the desired level of clinical resolution and defining the tasks for each problem in which the candidates were expected to be competent. After that came the determination that the OSCE was the most appropriate method of testing, considerations of issues of test administration, and selection of global rating as the method of scoring. With this in place, the investigator proceeded to examine the impact of juniors vs SPs as patients and raters on psychometric parameters in a senior candidate OSCE.

"Go forth into the world with courage; be strong, bold, and not afraid; Just Do It!" (Miller, 1993)

This initial work prompts deliberation about the use of junior students in clinical teaching and performance-based assessment. It is an exciting androgy. It is an opportunistic offering of qualitative learning benefits for both senior and junior participants. There is no doubt about the practicalities of using this rich resource in an educational program for health professionals: in financially challenged institutions faced with mounting a learning and assessment program using standardized patients, it is cost effective and efficient, affordable on a large scale in contrast to the use of professional actors. As a pool for standardized patients, juniors meet the requirements
of *availability* and *accessibility*. This research probed questions of *ability*. It sought to define areas for reflection and consideration, explore plausible applications which are operationally meaningful and suggest a template for useful implementation.

**Ability:** All health professional students should be taught and tested at some point before graduation using standardized patients. The use of junior students to facilitate this would appear to be as fair, acceptable and feasible as professional SPs; more economical; and as reliable and consistent as the SPs as reported in this analysis. There was no difference between the juniors and SPs tested in terms of the impact on faculty raters, patient raters and candidate performance. While valid across a smaller number of general performance measures, compared to SP or faculty raters, there was some evidence that when using JRs as patients, both faculty and junior scores are valid compared to measures which test process skills and may predict performance in future assessment of this nature.

The lower scores generated by TA raters as compared to either SP or JR raters supports other evidence of faculty as raters in the literature (Stillman, 1993) It reinforces the need to have faculty raters for high-stakes exams or those which there may be a high failure rate. However, using clinical faculty members as raters in standard OSCEs is also important since it forces the faculty to observe and evaluate the effectiveness of teaching, which they otherwise might not do. In contrast, these results can also be interpreted as important evidence to continue the use of SPs and JRs as raters, as they may clearly capture the evaluation of communication skills from the patient’s perspective.

**Areas for Reflection and Consideration:** While this study did not focus on examination of operational factors such as organization and training, this may be one area in which financial and time constraints are at odds. It is far simpler in terms of faculty time management to share or
delegate the training and scheduling demands of a small group of thirty professional actors with a professional trainer, if the financial budget supports this expenditure. In contrast, there were considerably increased demands on instructor time who was not assisted in any way, to deliver the same outcomes for the sixty students involved in the OSCE, including sorting the class cohort into the group selected for the exam, and training for consistency in role-portrayal and use of the rating tool. While students were accessible, careful scheduling around their timetable was needed to make training available at times which were convenient and reasonable in duration, yet close enough in proximity to the examination date to afford a measure of test security. There were other issues which required more careful planning in using of students vs actors. In consideration of their workload, the investigator was careful to schedule the examination date on a day which did not conflict too significantly with workload or examinations in other first year courses, and this was a weekend day. Students were not reimbursed for their participation and had to provide their own transportation which was problematic for some since the OSCE was not held during regular classes. Finally, in order to affect a high degree of realism in role-delivery, all junior cases were written to portray patients of this age group, whereas greater variability was possible with SPs, and the ages for their roles ranged from teenagers to the elderly. However, since the first year cohort was multi-cultural, it was possible to direct a variety of ethnicities into their roles, and this was not possible with the SPs, who were a much smaller cohort.

Although not part of the data analysis for this investigation, the use of the remaining first-year students for the practices also required the instructor's time for training and scheduling. However, this ensured a degree of fairness for the project, as all juniors were equally responsible for participating. Interestingly, when surveyed about these issues of time commitment, expense, convenience and a difference in the task requirements, the vast majority of juniors indicated their
overwhelming support of the project without complaint. For each cohort, the majority indicated they enjoyed their task, and would select it again if a choice were given, over other participation options (exam vs practices).

In this administration of the OSCE, SPs were used on day one (Saturday) and juniors were used on day two (Sunday). Feedback from juniors was useful in suggesting reversing this order, giving them an opportunity to use Sunday for relaxing, work or study for the coming week. Candidates made a similar suggestion, since they reported being more at ease with the juniors than the SPs during the interviews, and felt it would lessen their anxiety for SP interviews if they began with JRs. This suggestion made scheduling easier for the instructor and was implemented the following year.

Plausible Applications/Useful Implementation: Use of juniors for both practice and examination purposes occurred in this study to capture the full breadth and depth of applications which would benefit the seniors, and at the same time, offer an opportunity to each junior to participate and benefit in some way since the class was so large. However, changing this arrangement to exam only, or practice only, are both reasonable alternatives, depending upon time, class size, educational goals and finances. While juniors were not reimbursed for participating, there was a cost to the program for this examination day, in terms of reimbursement for faculty raters, invigilators, administration staff, and refreshments. If an operational budget would not support these expenditures, value to both teaching and assessment could still be achieved by using the juniors for formative practices. As the class size will increase for the Faculty of Pharmacy to 240 students in the next few years, use will be restructured to offer more practices or more test interviews for the OSCE, as required.
As enrollment increases, the financial burdens and time constraints could be offset by the introduction of sequential testing using JRs and SPs to best advantage. At the extreme, an initial screening examination could be held consisting of JRs as raters in a full test to excuse those with achievement above a cutoff level from undergoing the remaining examination. Those below the cutoff level would be examined in an OSCE with SPs. Alternatively, the initial screening and consequent OSCE could be composed of various ratios of JRs and SPs as economics and organization dictated. This would be a feasible way to control test costs and logistics. The literature would seem to support this in a preliminary way. Results of a retrospective analysis showed that performance on the first day (six cases) of a three-day examination (18 cases) was sufficiently predictive of performance on the full three-day examination that 56% would have passed the screening test, and the vast majority of those would have subsequently passed the full examination maintaining an acceptably low passing-error rate (0.4%) (Colliver et al, 1991)

More study is needed to validate the techniques and confirm the power of the method.

6.6 Recommendations for Future Research

6.6.1 Psychometric Questions

The challenge for clinical program development in the health care professions is to improve the quality of teaching and assessment while making it affordable, standardized, relevant and accessible to all. While this study was an important first step in the quantitative assessment of the use of JRs vs SPs in enhancing teaching and performance, the results are only
preliminary and the conclusions that can be drawn are small. It is important to continue to test the limits of applicability – in terms of psychometric parameters of reliability, reproducibility, validity and generalizability of this pedagogical approach. More evidence that can provide objective data is needed that using juniors as standardized patients, can substantially improve the quality of information for formative and summative evaluation of students. How might data from JR-based examinations complement SP-based examinations or written tests? The field of exploration is rich with further questions and the opportunities to examine them, especially considering the financial feasibility of using juniors, the large pool of subjects with increasing enrollment, and the requisites for performance-based assessment. Investigation of this nature would be part of the quality control of the curriculum.

Reliability: Further study of this OSCE is possible in future administrations where inquiry should consider the psychometric impact on reliability or consistency of a number of variations: reversing the sequencing of JRs vs SPs examination days, alternating use of JRs and SPs during one day, using SPs vs JRs to test the same cases, and increasing the number of stations or tests. An attempt could be made to decrease differences between JRs and SPs in preparation and delivery of their roles. The quality of training of JRs in role-portrayal could be enhanced through increasing training time or through the assistance of a professional trainer. The pool of JRs used in the OSCE could be made the same size as the SPs (thirty) so that each JR would also repeat their role eight times rather than four. The remaining students could be utilized in practices or as a second exam pool to add a third examination day and increase the number of interviews per candidate from four to six.

Validity: Concurrent validity could be further tested by comparing results with other selected performance measures not used in this study, which focus on processing, rather than
content skills. In May 2001, The Pharmacy Examining Board of Candidate held their first national SP-based OSCE for licensure qualification. The same cohort of candidates that participated in this study was participant in that examination. Comparison of results would enhance understanding of the *predictive validity* of this work. *Construct validation* is used to validate clinical competence by showing that expected relationships with other factors can be detected by the measurement process ie, clinical competence should increase with additional training. This was not tested in this study, but might be done in future investigations. Administering the OSCE to both second year and third year students, and perhaps even a test group of recent graduates of the program and faculty, may determine if the rating tool or type of patient used (JR vs SP) can better differentiate performance at different levels. Comparisons with experienced or expert clinical performance would add *predictive validity* evidence that relates performance in the OSCE to actual patient care outcomes.

*Generalizability:* Evidence of generalizability is also needed. Since the courses are studied consecutively over two years with the same instructor, it would be possible to test the same cohort of students at the both second and third year levels, and perhaps prior to graduation in fourth year, to see if the OSCE measures generalize across multiple administrations to the same students: are they stable over time and across content domains? Use of juniors as SPs would make this economically and practically feasible. It would be interesting also, to examine the impact of training as a patient and rater when a junior upon performance when a senior candidate. With such large classes, it might be possible to divide one junior cohort into various voluntary groups which would either not participate, participate in practices or in the OSCE in first year, and then compare performances in a third year OSCE.
Multicultural or gender variables: Because station scores depend on the judgment of observers, characteristics of raters or candidates unrelated to clinical skills could influence those evaluations. These characteristics include age, gender, ethnicity, accent, and physical appearance. The large multicultural cohort of students at the Faculty which have comparable male/female ratios provides a good subject pool to study racial, ethnic, cultural and gender factors affecting performance in clinical skills testing. Psychometric testing could be done comparing groups of both JRs and candidates stratified into contrasting subgroups in the variable being examined. Combining use of JRs with SPs would also provide the opportunity for quantitative testing of the effects of patient age as a variable affecting performance testing, allowing patient test groups of young adults to be compared to teenagers, middle-aged or the elderly. Such bias studies may be especially important if SP-based tests are to be used for assessing the clinical skills of foreign graduates. Some evidence in the literature exists that the interaction between student and SP gender has no effect on student scores and ratios (Furman et al. 1993; Gispert et al. 1999), and that there is no indication of an examinee- by- SP ethnicity interaction (Colliver et al. 1999) However, cultural competency deficits and differences were measurable in an SP OSCE health-beliefs station. (Robins et al. 2001) Further study is needed to support or expand these areas of investigation.

6.6.2 Qualitative Questions

"...the problem seems in part that we dichotomize (SP) use into teaching and evaluation. It might be more helpful to acknowledge that this is not a dichotomy but actually a continuum. since evaluation is certainly the most powerful teaching tool we possess. While a curriculum defines what will be taught, it is evaluation that dictates what must be learned." (Miller, 1993)
It is useful to reflect upon where the use of first year students as standardized patients fits into this continuum by considering what the junior participant is able to do. The first is demonstration: juniors can be used as standardized patient instructors, to demonstrate something difficult or uncomfortable to learn from real patients. Beyond that, junior standardized patients can be employed in the summative evaluation of learning, to judge whether educational goals have been achieved. Because of their judgmental character, the psychometric properties of validity and reliability are of central importance here. Finally, junior students can facilitate learning through formative evaluation exclusively for the purpose of feedback. Candidates who undergo training practices with juniors can be assured of a non-judgmental assessment that allows them to reveal what they do not know for the purposes of learning without fear they will be judged unfavourably. For this purpose, the psychometric qualities of the test are relatively unimportant.

The behaviours a pharmacist, or others health care professionals, must learn are an integration of a complex set of components illustrated in the familiar diagram shown in Figure 4. The well-informed professional must know how to use the knowledge he possesses. Nevertheless, readiness for practice is determined by the ability to demonstrate that he is not only skillful at processing this knowledge during observation and assessment but can carry out appropriate action independently. It is at the upper two levels of this
pyramid that juniors are most useful as standardized patients. Both the OSCE itself and the formative practices are arenas in which the junior student can enhance the learning of the candidate through feedback and assessment. The private individual practices between junior and senior students before the OSCE are a way to enable autonomous action in the absence of any observation.

Feedback from the candidates reflected their positive feelings about the qualitative value of this experience. They commented on the increased gains in the professional practice, clinical problem-solving and communication skills. They were unanimous in their support of the continuation of this program as a facilitative way to acquire professional competencies and generated and unanimously signed a petition to ensure it would be mandatory for future classes.

"I hear and I forget
I see and I remember
I do and I understand"
(Anonymous)

But equally important to consider are the learning gains for the junior students who participated in this study. While a few preliminary studies in the literature review examined the subjective experiences of being a standardized patient, the impact on junior students should be considered in terms of advantages which may be long-term.

"We cross borders, but we don't erase them, we take our borders with us." (Behar, 1993)

The junior students in the study were encouraged to recognize that their subjectivity was engaged purposefully, and to use this lens to inquire into perspectives and interpretations, to
shape new questions through examining assumptions, and to explore feelings in relationship to learning.

Juniors, either in practice sessions or after the OSCE, completed qualitative feedback surveys which asked them to describe their experience, reflect on how they felt during the situation, analyze why this was so, and suggest changes for future implementation. Juniors reported gains in many areas: ability to experience peer-teaching, ability to network with senior students, training for their own future OSCE, increased tolerance, a balanced health care perspective in terms of the patient's views and needs, increased communication skills, and introduction to the long-term goals of the program. They felt empowered by the ability to effect program development through the use of formative feedback. Through the direction of the feedback form, they acquired a reflective practice that encourages them to acknowledge, embrace, challenge or augment gains in scholarship.

"Learning is discovery" (Miller, 1993)

There has been much rhetoric but little empirical work on the educational impact of SP-based tests, aside from Newble's efforts at the University of Adelaide. He investigated the impact of an SP OSCE on student study habits, through questionnaires completed before and after the exam. Results indicated that the exam had a dramatic impact on how students spent their time, decreasing studying and increasing skill preparation, and students indicated a high level of satisfaction with the exam and assertions of its relevance in preference to written testing. (Newble, 1988) More research in this area is needed, because the hypothesized educational
impact of SP-based tests has been a major factor in their increased use, despite high costs and psychometric shortcomings.

The similarly encouraging perceptions from student participants in this study OSCE about the positive educational impact reflect the perspectives of the single cohort of junior and senior students used. They were gathered in order to identify if there were perceived benefits worthy of further investigation. It is important to explore these preliminary indications that there may be significant qualitative benefits to both junior and senior learners when using first year students as standardized patients in training and administering an OSCE. Many questions will need to be answered before this training program can be fully developed. Can the rewards for participants be clearly identified and projected? How can these rewards be enhanced by the behavior of faculty and students? What further training is necessary to deliver these goals with sufficient efficiency and effectiveness? Can the learning gains be extended to other programs in other disciplines? Are they transferable to other providers of professional education?

In the future, this investigator would like to assess the acceptability of using first year students as standardized patients for a senior candidate OSCE and to study the learning benefits to student participants more intensively. Does the design of the program affect the way either group approaches learning? If so, what strategies do they use and how are they different? How could they be facilitated by faculty? What are the reflections and opinions of faculty raters on the impact of this program?
"Good ethnography is an intellectual exorcism in which, forced to take the perspective of the other, we are wrenched out of our self." (Sweder, 1986)

An appropriate approach might be an ethnographic study. These questions might be addressed through triangulation and the use of surveys, focus groups and individual participants. The information could then be analyzed, interpreted and critically appraised. Interventions of countermeasures and adjustments could be designed to expand the breadth and depth of each learning cycle of research into a sound database of educational knowledge, theory and insights.

6.7 Final Comments

"The Road goes ever on and on
Down from the door where it began
Now far ahead the Road has gone,
And I must follow if I can.
Pursuing it with weary feet,
Until it joins some larger way,
Where many paths and errands meet.
And whither then? I cannot say." (Tolkien, 1965)

It is hoped that this study will serve as a catalyst for continued research on the use of junior students as standardized patients. To give form and meaning to these results and conclusions, the perspective must shift from the specific situation in this study to one which is more global. Additional research is indicated in order to establish the trustworthiness of these data interpretations and achieve generalizability, projections, insights and questions for further direction, which could be disseminated across health professional education. This evidence must
be a balance of quantitative and qualitative considerations in order to effect a change of paradigm. Acceptance of new methodology is based not just on objective data, but on attitudes.

Can such investigations lead to a shared view on the value and role of networking between junior and senior students in education and evaluation? Can they be extrapolated across disciplines and across programs? Can they be refined to transform the student into a professional who has not only achieved expertise in all the required competencies, but who is a reflective, life-long, self-directed learner? This research is the first step in a path that may lead to enhanced outcomes for both health care professionals and the patients they serve.
References


Annex to the proceedings of the AAMC consensus conference on the use of standardized patients in the teaching and evaluation of clinical skills, Teach Learn Med 6 (1994): 2-35

Austin, Z: Use of compressed cases, an extended family tree, and simulated patient-actors to teach pharmacy practice skills. Poster Presentation at American Association of Colleges of Pharmacy Annual General Meeting, July 2000, San Diego, CA.

Baerheim A and Malterud K Simulated patients for the practical examination of medical students: intentions, procedures and experiences Medical Education 29 (1995) 410-413

Barnes HV, Albanese M, Schroeder J, Reiter S, Senior medical students as patient-preceptors to introduce basic history and physical examination skills to second year medical students Annu Conf Res Med Educ 16 (1977):51-6


Barrow HS. Simulated patients The Development and Use of a New Technique in Medical Education Charles C Thomas, Springfield, Illinois. 1971


Barrows HS Peter, MJ. (Eds) How to Begin reforming the Medical curriculum Springfield, Illinois: Wouthern Illinois University School of Medicine, 1984


Belanger A, There are ‘continuing education alternatives’ to alternative continuing education’ in physiotherapy: President’s report Physiotherapy Canada 49(4) (1997): 249-252


Chang EH and Power DV, Are medical students comfortable with practicing physical examinations on each other? Academic Medicine 75(2000): 384-9


Cohen DS, Colliver JA, Marcy MS, Fried ED, Swartz MH, Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills Acad Med 71 (1996) S87-89


Collis KF, Biggs JB Classroom examples of cognitive development phenomena: the SOLO taxonomy Australia (1979): 217 pages

Colliver JA et al, Three studies of the effect of multiple standardized patients on inter-case reliability of five standardized patient examinations Teach Learn Med 2 (1990): 237-245


Colliver JA et al, Test security in standardized patient examinations: Analysis with scores on working diagnosis and final diagnosis In Proceeding of the Thirtieth-first annual conference on Research in Medical Education Acad, Med 67 Supplement (1992) S7-S9


Colliver JA, Marcy ML, Vu V, Stewart, DE, Robbs RS, The effect of using multiple standardized patients to rate interpersonal and communication skills on the same case on the inter-case reliability of the ratings. Paper presented at the annual meeting of the American Educational Research Association, Division I, Atlanta, Georgia, April, 1993


examination administer over an extended testing period Academic Medicine 73 October S (1998): S81-S83


Colliver JA, Vu V, Marcy ML Robbs RS, Effect of examinee gender, and their interaction on standardized patients’ ratings of examinees’ interpersonal and communication skills. Acad Med 68 (1993) 153-7

Cooper C and Mira M, Who should assess medical students’ communication skills: their academic teachers or their patients? Med Education 32 (1998) 419-421


Davis M, The way ahead: Teaching with simulated patients Medical Teacher 11 (1989): 319


De Champlain AF, MacMillan M, King AM, Klass DJ, Margolis MJ, Assessing the impact of intra-site and inter-site checklist recording discrepancies on the reliability of scores obtained in a nationally administered standardized patient examination. Academic Medicine 74 (10 supplement) (1999): S52-S54


Ebel R, Must all tests be valid? Amer Psych 16 (1961): 640-7


Frederiksen N, the real test bias: Influences of testing on teaching and learning Amer Psych 39 (1984): 193-202


Gispert R, Rue M, Roma F, Martinez-Carretero FM. Gender, sequence of cases and day effects on clinical skills assessment with standardized patients Medical Education 33(1999): 499-503


Harden RM Assess clinical competence – An overview, Medical Teacher 1 (1979): 289-296


Hawkins R, Gross R, Beuttel S, Hombe E, Use of Standardized Patients for Teaching and Evaluating the Genitourinary Examination Skills of Internal Medicine Residents. Teaching and Learning in Medicine, 10(2) (1997): 65-68

Helfer RE, Black MA & Teitelbaum H, A comparison of pediatric interviewing skills using real and simulated mothers Pediatrics 55 (1975) 397-400


Hodges B, Tiberius R, Hanson M, Regehr G McNaughton N. OSCE checklists do not capture increasing levels of expertise. Paper presented at the Annual Meetings of the Association of American Medical Colleges, New Orleans LA, November 1998


Houle CO Continuing Learning in the Professions San Francisco: Jossey-Bass Publishers 1980


Johnson SM, Kurtz, ME, Tomlinson, R Howe, KR. Students' Stereotypes of Patients as barriers to Clinical Decision-making. J Medical Education. 61 (1986):727-737


King AM, Perkokowski-Roger L, Pohl S. Planning standardized patient programs: Case development, patient training and costs. Teaching and Learning in Medicine 6 (1994): 6-14


Kopp KC and Johnson JA. Checklist agreement between standardized patients and faculty. J Dental Education 59 (1995): 824-9


Lamont CT & Hennen BKE. The use of simulated patients in a certification examination in family medicine. Journal of Medical Education 47 (1972) 789-792


Mann KV, MacDonald AC, Norcini JJ. Reliability of objective structured clinical examinations: Four years of experience in a surgical clerkship Teaching and Learning in Medicine 2(4) (1990): 219-224


McAvoy BR. Teaching clinical skills to medical students: The use of simulated patients and videotaping in general practice Medical Education 22(1988):193-99


Meadow R & Hewitt C. Teaching communication skills with the help of actresses and video-tape simulations. British Journal of Medical Education 6 (1972) 317-322


Miles MB. Huberman AM. Qualitative data analysis: An expanded sourcebook. Thousand Oaks: Sage Publications, 1994


Nelson, ITP Canadian Dictionary of English Language Toronto, Ontario International Thomson Publishing Company


Newble D, Eight years’ experience with a structured clinical examination. Medical Education 22 (1988):200-4,


Norcini JJ, Recertification in medical specialties Academic Medicine 69 (supple) (1994): 90-4


Osler W. The hospital as a college In Aequanimitas, and Other Addresses, Philadelphia. Blakiston’s, (1904) 329-42.


Papadakis MA, Croughan-Minihane, M. Fromm LJ, Wilkie, HA, and Ernester, VL, A comparison of two methods to teach smoking-cessation techniques to medical students


Sanson-Fisher RW and Poole Ad, Simulated patients and the assessment of medical students’ interpersonal skills Medical Education 14 (1980): 249-53


Sloan DA, Donnelly MB, Johnson SB et al, Use of an objective structured clinical examination to measure improvement in clinical competence during the surgical internship. Surgery 114 (1993): 343-351


Stillman P, Swanson D, Sme S et al. Psychometric characteristics of standardized patients for assessment of clinical skills (Final Report to the American Board of Internal Medicine), 1986


Swanson DB, Norcini JJ Factors influencing the reproducibility of tests using standardized patients Teach Learn Med 1 (1989): 158-166


Tamblyn RM, Klass DJ, Schnabl GK, Kopelow ML, Multiple station examinations: Noise in the system Acad Med 165 (1990): 555-6

Tamblyn RM, Klass DJ, Schnabl GK, Kopelow ML, Sources of unreliability and bias in standardized patient rating. Teach Learn Med (3) 1991: 74-85


Thomas RE, Teaching medicine with cases: student and teacher opinion. Medical Education 26 (1992): 200-207


Tolkien JRR, The Fellowship of the Ring, Houghton Mifflin Co. 1965

Traub RE Facing the challenge of multidimensionality in performance assessment. In Rothman AI, Cohen R (Eds) Plenary Presentation at the Sixth Ottawa Conference on Medical Education. Toronto, June 1994


Usherwood T, Subjective and behavioural evaluation of the teaching of patient interview skills. Medical Education 27 (1993):41-47


Van der Vleuten CPM, van Luyk S, Ballegooijen A, Swanson D, Training and experience of medical examiners Medical Education 23(1989): 290-6

Van der Vleuten CPM and Swanson DB Assessment of clinical skills with standardized patients: State of the art, Teaching and Learning in Medicine 2(1990): 58-76

Vu NV, Barrows HS, Use of standardized patients in clinical assessments: recent developments and measurement finding Educational Researcher 23 (1994): 23 -30


Appendices
Appendix One: Summary of Instrumentation (Data Analysis) to be Used

1. **Reliability: Is there an effect on inter-station correlations for the following dependent measures?**
   a. Correlation of TA scores between station 1 and station 2 when using SPs
   b. Correlation of TA scores between station 1 and station 2 when using JRs

   Use an OLIKINS Z test to compare the two correlations

2. **Reliability: Is there a systematic influence on candidate scores when the means are compared?**
   a. Mean of TA scores on station 1 and station 2 when using SPs
   b. Mean of TA scores on station 1 and station 2 when using JRs

   Use a paired t test to compare the two correlations

3. **Concurrent Validity of Scores: Compare the GRS scores with other measures:**
   a. Comparing the average TA scores using SPs vs other measures
   b. Comparing the average TA scores using JRs vs other measures

4. **Reliability: Is there an effect on inter-station correlations for the following dependent measures?**
   a. Correlation of SP scores between station 1 and station 2
   b. Correlation of IR scores between station 1 and station 2

   Use an OLIKINS Z test to compare the two correlations

5. **Reliability: Is there a systematic influence on student scores when the means are compared?**
   a. Mean of SP scores between station 1 and station 2
   b. Mean of IR scores between station 1 and station 2

   Use a paired t test to compare the two means

6. **Concurrent Validity of Scores: Compare the GRS scores with other measures:**
   a. Comparing the SP scores vs other measures
   b. Comparing the IR scores vs other measures

7. **An additional analysis for concurrent validity can be made which compares:**
   a. The correlation of SP vs TA scores for station 1 and station 2
   b. The correlation of IR vs TA scores for station 1 and station 2

   Use an OLIKINS Z test to compare the two correlations
Appendix 2: Case Template

Anticipated Time needed (mins): 10 minutes

Student tasks (check all that are appropriate):
- Deal with a communication issue
- Obtain focused and relevant history
- Deal with an emotional issue
- Deal with an ethical issue
- Identify and resolve or prevent ______ drug-related problems
- Provide patient education
- Counsel the caregiver for patient or the patient
- Discuss monitoring and followup

Main Drug Related Problems:

1. 
2. 
3. 

1. Case Summary

2. Description of the Case:

   A. **Type of encounter:** community pharmacy visit
   B. **Location/setting of encounter:** Community Dispensary
   C. **Opening statement the patient will say to the pharmacist:**
   D. **Description of Patient:**

   Name:

   Age:

   Gender:

   Weight:
Socioeconomic and education level:

Language spoken:

Appearance:

Dress:

E. Patient Behaviour, affect, mannerisms, non-verbal behaviour,

F. Patient’s agenda:

G. Any questions the patient will consistently ask: what she should do exactly regarding any treatment suggestions

3. Patient history:

A. Chief complaint:

B. Current medical problems/medications:

C. Medication List:

D. Relevant past medical history

E. Allergies:

F. Immunizations:

G. Relevant Social History

H. Relevant Family History:

I. Compliance Issues:

J. Drug Related Problems:
   - see page 1

4. Communication Challenges Encountered and Expected Behaviour of Candidate:
5. Ethical Issues Raised:

6. Critical Issues (Key Points):
Supply the main 'must-have' issues below and how the student should resolve each one in the Checklist provided as attached pages
1.
2.
3.
4.
5.

7. Props Needed:

8. Additional Information to help explain case to the Standardized Patient:

9. Short Introduction of the case for the Patient:
Give the opening line for the patient, with the appropriate emotional tone:

This standardized patient form was developed by Lisa Dolovich of the Faculty of Pharm. Phm D program and amended for use in 220/320 by Debra Sibbald.
Appendix Three: OSCE Global Rating Scale

Student __________________________  Case # _______  Time _______  Examiner _______

<table>
<thead>
<tr>
<th>OVERALL ASSESSMENT OF THE KNOWLEDGE AND SKILLS DEMONSTRATED IN THE INTERVIEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>Responds inappropriately and ineffectively to the task indicating a lack of knowledge and/or undeveloped interpersonal and interviewing skills.</td>
</tr>
</tbody>
</table>

Circle the rating which best reflects your judgement of the student's performance in the following categories:

RESPONSE TO PATIENT’S FEELINGS AND NEEDS (EMPATHY)

<table>
<thead>
<tr>
<th>RESPONSE TO PATIENT’S FEELINGS AND NEEDS (EMPATHY)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>Does not respond to obvious patient clues and/or responds inappropriately.</td>
</tr>
</tbody>
</table>

DEGREE OF COHERENCE IN THE INTERVIEW

<table>
<thead>
<tr>
<th>DEGREE OF COHERENCE IN THE INTERVIEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>No recognizable plan to the interaction, the plan does not demonstrate cohesion, or the patient must determine direction of the interview.</td>
</tr>
</tbody>
</table>

VERBAL EXPRESSION

<table>
<thead>
<tr>
<th>VERBAL EXPRESSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>Communicates in manner that interferes with and/or prevents understanding by patient.</td>
</tr>
</tbody>
</table>

NON-VERBAL EXPRESSION

<table>
<thead>
<tr>
<th>NON-VERBAL EXPRESSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>Fails to engage, frustrates and/or antagonizes the patient.</td>
</tr>
</tbody>
</table>

Key Points

<table>
<thead>
<tr>
<th>Key Points</th>
<th>Yes</th>
<th>No</th>
<th>Comments:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>