AN EVALUATION OF THE UTILITY OF A HYBRID OBJECTIVE STRUCTURED CLINICAL EXAMINATION FOR THE USE OF ASSESSING RESIDENTS ENROLLED IN McMASTER UNIVERSITY'S ORTHOPAEDIC SURGERY RESIDENCY PROGRAM

AN EVALUATION OF THE UTILITY OF A HYBRID OBJECTIVE STRUCTURED CLINICAL EXAMINATION FOR THE USE OF ASSESSING RESIDENTS ENROLLED IN McMASTER UNIVERSITY'S ORTHOPAEDIC SURGERY RESIDENCY PROGRAM

VANJA GAVRANIC, BHSc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCES (2013) Hamilton, Ontario (Health Research Methodology)

TITLE: An Evaluation of the Utility of a Hybrid Objective Structured Clinical Examination for the use of Assessing Residents Enrolled in McMaster University's Orthopaedic Surgery Residency Program AUTHOR: Vanja Gavranic, BHSc (McMaster University) SUPERVISOR: Kelly L Dore, PhD (McMaster University) NUMBER OF PAGES: 115

Abstract

Introduction: McMaster University's orthopaedic surgery residency program implemented the OSCE as an assessment tool in 2010; this study evaluates the first four OSCEs administered to residents. The OSCEs were composed of knowledge-testing stations, which are normally not included in this testing format, and performance-testing stations. Recruiting enough faculty evaluators challenged the ability to feasibly implement this examination format. Knowledge-testing stations were incorporated since they do not require evaluators to be present. Reliability was assessed, and the correlation between knowledge-testing station scores and performance-testing station scores was determined. The ability of the OSCE to discriminate between residents in different post-graduate years (PGYs) was assessed. Residents' acceptability of the OSCE was also assessed. **Methods:** Reliability was assessed using generalizability theory. The correlation of knowledge-testing and performance-testing station scores was measured with Pearson's r. A two-way ANOVA was used to analyze whether the OSCE can discriminate between residents in different PGYs. An exit survey was administrated after each OSCE to assess acceptability. **Results:** The generalizability estimates of each OSCE ranged from 0.71 to 0.87. The disattenuated correlation between knowledge- and performance-testing stations for senior residents was 1.00, and 0.89 for junior residents. A significant effect of year of residency was found for the October 2010 OSCE in the ANOVA (F(1,30) = 11.027, p = 0.005), but the remaining OSCEs did not replicate this finding. In general, residents felt that they were able to present an accurate portrayal of themselves in the OSCEs and that the examination covered appropriate topics.

Discussion: The OSCEs were reliable and acceptable to residents. The high correlations between knowledge- and performance-testing station scores suggest that the examination can be made more feasible by including knowledge-testing stations. The small sample sizes made significant differences difficult to detect between levels of training, resulting in inconclusive evidence for this construct validation measure.

Acknowledgements

Thank you to my supervisor, Kelly L Dore, for all of her help with this project, and for being such a wonderful and supportive mentor. Thank you to everyone at PERD for the wonderful experience it has been to work with you over the past two years, and especially thank you to Mahan Kulasegaram at PERD, who has additionally been a wonderful help to me in my graduate work, including this thesis project. I would also like to thank Geoffrey R Norman for his guidance in this project.

I would like to thank Bradley A Petrisor and my committee members Michelle A Ghert and Mohit Bhandari for their helpful comments during this project.

I would also like to thank my former professor Stash Nastos for all of his help throughout my undergraduate program and graduate school.

In my personal life, I would like to thank those who have helped me not only with getting through graduate school, but who are a constant source of love and support: my parents Rajko and Jela Gavranic, my brother Srdjan Gavranic, my cousin Tamara Erak, my friends Marija Vukmirovic and Rados Panic, and my boyfriend Paul Kuyanov.

Table of Contents

List of Figures and Tables

List of all Abbreviations and Symbols

OSCE, objective structured clinical examination

MMI, multiple mini interview

RCPSC, Royal College of Physicians and Surgeons of Canada

PGY, post-graduate year

MCQ, multiple choice question

ITER, in-training evaluation report

SAQ, short-answer question

FITER, final in-training evaluation report

MCCQEI, Medical Council of Canada Qualifying Examination Part I

PMP, patient management problem

G, generalizability

ICC, intraclass correlation coefficient

P, resident

Y, year of training

S, station, or rater and station

:, nested

×, crossed

τ, variance from facet of differentiation

δ, error variance

SD, standard deviation

Chapter 1: Introduction

1.1 McMaster University's Orthopaedic Surgery Residency Program

The Division of Orthopaedic Surgery at McMaster University is a division within the Department of Surgery, under the Faculty of Health Sciences. Each year, approximately six new residents are admitted into the five-year training program.

Applications to the residency program are managed through the Canadian Residency Matching Service. Successful admission into the residency program is based on a final ranking of applicants based on the quality of their written application and interview. Since 2010, the interviews have been conducted in the multiple mini interview (MMI) format (Dore et al., 2010). After the interviews are complete, the faculty members rank the applicants that have demonstrated to be a good fit for the program. Entrance into the program became stringent and competitive in 2010, in order that students who have not demonstrated that the potential to be hardworking residents with an understanding of the field of orthopaedics are not accepted.

The Royal College of Physicians and Surgeons of Canada (RCPSC) has outlined the minimum training requirements for residents throughout their time in an orthopaedic surgery residency program. These requirements are outlined in Appendix 1, and include rotations in orthopaedic surgery and its subspecialties, critical care, a service that provides trauma management (e.g., emergency medicine), general and/or vascular surgery, and internal medicine (Royal College of Physicians and Surgeons of Canada, Specialty Training Requirements in Orthopaedic Surgery, 2013).

Residents spend their first two (junior) program years (PGYs) completing training in basic surgical principles. Much of this education comes from clinical rotations in orthopaedic surgery and general surgery. Junior residents also complete rotations in emergency medicine and internal medicine. In the senior years (PGY 3, PGY 4, and PGY 5), residents complete rotations in eight orthopaedic surgical subspecialties: trauma, paediatric orthopaedics, foot and ankle, major joint reconstruction, spine, upper extremity, pathology/oncology, and sports medicine.

Throughout the five years in the program, residents also attend educational lectures, which include morbidity and mortality talks, orthopaedic lectures on CanMeds, grand round presentations, senior session presentations, paediatric presentations, general lectures, ortho "must haves", oncology lectures, and evidence-based orthopaedics lectures.

The aim of the residency training is to produce competent orthopaedic surgeons, in accordance with the expectations outlined by seven domains of competence in the RCPSC CanMeds framework: medical expert, scholar, professional, communicator, collaborator, manager and health advocate (Royal College, 2011).

Residents are expected to demonstrate and maintain these competencies in a field that is continuously changing and therefore placing greater demands on the level of expertise expected of them. They are accountable to both patients and peers, and quality assurance of trainees' performance using a multitude of assessment tools is crucial to assuring they are meeting the level of expertise expected of them.

1.2 Assessment in McMaster University's Orthopaedic Surgery Residency Program

1.2.1 Introduction to Assessment

Placing the evaluation of McMaster University's orthopaedic surgery residents into the appropriate context warrants a discussion on assessment in general and current trends in the assessment of postgraduate medical trainees.

Historically, training programs have used knowledge-testing examinations such as multiple-choice questions (MCQs) to assess medical students and residents (Norman, 2002). Additionally, licensing boards have exclusively used such tests for certification examinations (Reznick et al., 1992). The psychometric properties of these tests have been studied extensively, and it is shown that they can produce reliable and valid estimates of clinical competence (Neufeld, 1985).

During the latter half of the twentieth century, assessment in medical schools and residency programs shifted away from exclusively the evaluation of knowledge, and placed a greater emphasis on the direct observation of medical trainees (Howley, 2004). Licensing bodies also began to assess performance in certification examinations. For example, since 1982, the Medical Council of Canada has been using the objective structured clinical examination (OSCE), a measure of performance, as the final part of the testing process for medical graduates to obtain their license to practice medicine (Reznick et al., 1992). The shift towards the inclusion of performance-based assessment was based on the belief that measuring knowledge of medical trainees, although essential, would not be sufficient to assess their ability to demonstrate their competence, especially in a clinical setting (Miller, 1990).

The distinction between different levels of assessment was made by George E. Miller, using a framework commonly known as Miller's pyramid (1990), shown in Figure 1 in Appendix 2 (Miller, 1990). Knowledge, placed on the bottom of the pyramid, is often measured through factual tests, such as MCQs (Wass et al., 2001). The next level of assessment, competence, is defined as the ability of a medical professional to apply their knowledge. For example, competence may be assessed by the ability of a medical professional to gather information, interpret the information, and incorporate it into a management plan (Miller, 1990). This is tested with the same testing formats used to assess knowledge (Wass et al., 2001). Performance, the ability of medical trainees to demonstrate that they are competent, is the next level of the hierarchy. Performance of trainees is evaluated by placing them into hypothetical scenarios, sometimes using standardized patients, and allowing them to demonstrate how they would approach those scenarios (Wass et al., 2001). An OSCE, which will be described in section 1.2.2 of this chapter, is commonly used to assess performance. At the top of the pyramid is action; that is, what the medical trainee or professional actually does in a clinical setting. A common method of assessing action is the in-training evaluation report (ITER), in which evaluators rate trainees in a variety of domains, such as the CanMEDS competencies. Although it would be ideal to evaluate trainees' performance in a clinical setting, methods such as the ITER have several limitations, such as poor reliability and a tendency for the evaluations to be completed weeks after students complete a rotation (Feldman et al., 2004; Finlay et al., 2006, Turnbull and Barneveld, 2002). This has led to the increasing popularity of using performance testing methods, in particular the OSCE, to predict action.

In spite of the trend towards performance-based and action-based assessment, it has been shown that the results of performance-testing and knowledge-testing examinations correlate well. For example, Matsell et al. (1991) and van Dalen et al. (2002) found the correlations between the two testing formats to be around 0.60 (Pearson's r). This indicates that the results of knowledge tests can be predictive of the results of performance tests and vice versa. Although the correlation between these testing formats is reasonably high, they do not actually measure the same constructs; that is, the knowledge could be seen as foundational to the performance. The behavioural aspects of clinical-problem solving are only directly observed in performance-based tests (Wass et al., 2001).

Assessment can be either formative or summative. Formative methods of assessment allow students to gauge their progress through a curriculum, and are used as benchmarks to allow students to evaluate their progress (Epstein, 2007). They allow students to use the assessment as a learning tool that provides feedback on their performance, which they may use to improve (Wass et al., 2001). On the other hand, feedback is not the primary goal of summative assessment; rather, the assessment is used to make a conclusion about the student's competence (Wass et al., 2001). It may be used in decision-making, such as in deciding whether a student may further their training or begin to practice (Epstein, 2007).

1.2.2 Methods Used to Assess Residents in McMaster University's Orthopaedic Surgery Residency Program

The assessment of residents in McMaster University's orthopaedic surgery residency program encompasses all four levels of Miller's pyramid. The program evaluates residents based on all seven CanMeds competencies.

The RCPSC requires the assessment of action with the ITER in order for a residency program to maintain its accreditation. An ITER is completed for each resident at the end of each rotation for all residents in McMaster University's orthopaedic surgery residency program. Each ITER is composed of seven subheadings that reflect the CanMeds domains of competence, and each subheading has a variable number of scales to rate different aspects of that domain of competence. In addition to ITERs, the orthopaedic surgery residency program also measures action with clinical encounter cards, clinical evaluation cards, operative evaluation cards, and patient surveys. These are all designed following a similar structure to the ITER (i.e. various domains of competence are assessed with a global rating scale at a variable time after a clinical interaction).

McMaster University's Orthopaedic surgery residency program has recently started to measure performance using the OSCE format. An OSCE is an examination in which examinees are required to rotate through multiple stations in which they are required to demonstrate a variety of skills. For example, they may be required to take the history of a standardized patient or perform a simulation of a procedure. Stems, descriptions of what the examinees are required to do at a particular station, are posted on the door of each station for examines to review prior to entering. An example of a station stem could be, "When you enter the room, you will be required to take the history of the patient who was admitted to the emergency room two hours earlier with severe calf pain." Examinees are given a limited amount of time to read the stem, after which a bell rings to signify that they must enter the station. After the allotted time for the station has passed, a bell rings to inform examinees to leave that station and move to the next one. There are evaluators at each station who rate examinees on their performance. Although examinees rotate through stations, faculty evaluators remain in the same room, allowing the evaluators to observe multiple students' performance on the same station. Examinees sign confidentiality statements to ensure that they will not share the station content.

The OSCEs conducted in McMaster University's orthopaedic surgery residency program are composed of knowledge-testing stations (MCQs, short answer questions (SAQs), and radiographs to diagnose) and performance-testing stations (residents are expected to discuss their approach to the management of a hypothetical case or to discuss a CanMeds competency).

The orthopaedic in-training examination is a multiple choice test, administered annually by the American Academy of Orthopaedics. This is an assessment of knowledge and competence. Senior residents also complete multiple choice examinations at the end of each teaching block. As mentioned, knowledge and competence are also tested in OSCE stations that require residents to answer various MCQs or SAQs or to make diagnoses using radiographs.

ITERs are the only forms of summative assessment for residents in each rotation. The final ITER (FITER) is an additional form of summative assessment in the program. It determines whether residents may complete the RCPSC Comprehensive Objective Examination. The FITER is completed by the residency program director, and asks whether a resident has acquired the competencies outlined by the RCPSC and is competent to practice, and which evaluations of the resident were taken into consideration in making this decision (i.e., written exams, clinical observations (e.g., ITERs), feedback from health care professionals, completion of a scholarly project, oral exams, OSCEs, and other evaluations).The FITER also presents rating scales for each domain of competence as outlined by CanMeds, and asks that the program director answers on an adjectival scale the extent to which the resident has met expectations in various aspects of each domain of competence. All other methods of assessment used in the program are formative.

1.3 Principles of Assessment – Utility Theory

Ideally, if a method of assessment is to be used to evaluate learners, it should have evidence to support its intended use in a particular setting and the intended interpretation of the scores. In 1996, Van der Vleuten proposed utility theory, a set of five criteria that should ideally be met in order to support the use of an assessment tool. In this utility framework, he outlined five key principles that must be met:

1. Reliability;
2. Validity;
3. Acceptability;
4. Feasibility; and
5. Educational impact (Van der Vleuten et al., 1996).

Reliability is defined as the proportion of total variance in scores that is due to the variance contributed by those being assessed (Streiner and Norman, 2008). Error contributes to the total variance; the larger the error the greater the total variance, and thus the smaller the proportion of variance due to differences in those being assessed (Streiner and Norman, 2008). In order that the proportion of variance contributed by differences between those being assessed by the measurement tool is high, the measurement tool must be able to discriminate between those being assessed (i.e., if everyone was given the same score, the proportion of variance due to differences between those being assessed would be 0 and thus reliability would be 0). Thus, discriminatory power and consistency are the key components of reliability.

Validity is defined as the evidence generated to support a particular interpretation of scores (Cook and Beckman, 2006). The process of validation is thus a process of testing hypotheses in order to support a particular interpretation of scores (Cook and Beckman, 2006). There are five different sources of evidence that can be collected to support construct validity: content, response process, internal structure, relationships to other variables, and consequences (Cook and Beckman, 2006). “Content” refers to the appropriateness of the items selected for the test, and whether they are representative of all domains of the construct being assessed (Cook and Beckman, 2006). “Response process” refers to the data-gathering process; that is, whether data have been entered accurately, whether data were combined appropriately to generate a composite score, and whether the appropriate method of scoring was used (e.g., checklists versus global ratings) (Downing and Haladyna, 2009). “Internal structure” refers to the expectation that scores intending to measure one construct should produce homogeneous results whereas scores intending to measure multiple constructs should produce heterogeneous results in a pattern that is predicted by the construct (Cook and Beckman, 2006). “Internal structure” refers to the expectation that there should be systematic differences between scores generated from different subgroups (such as residents at different levels of training in an OSCE) when the construct is measured (Cook and Beckman, 2006). “Relations to other variables” refers to the expectation that scores generated with the assessment tool should correlate with other methods of assessment that are intended to measure the same construct (Cook and Beckman, 2006). “Consequences” refers to the intended and unintended consequences of the use of the assessment tool (Downing and Haladyna, 2009).

Acceptablity refers to the extent to which faculty and students support the use of the assessment tool (Van der Vleuten, 1996).

Feasibility refers to the cost of an examination (Van der Vleuten, 1996). It also refers to other logistical challenges that may need to be overcome in order that the assessment tool may be implemented, such as the availability of faculty, staff, trainees, space, and any equipment that may be required.

The assessment of educational impact informs the “consequences” component of modern validity theory. Assessment drives learning, in terms of content, format, and programming (i.e. frequency of assessment) (Van der Vleuten, 1996). Evidence supporting educational impact should discuss or demonstrate that assessment has, in fact, driven learning (e.g., students practice hypothetical case management scenarios in preparation for the OSCE, students comment that the OSCE has identified limitations in their knowledge they intend to remedy).

1.4 Rationale for Introduction of the OSCE as an Assessment Tool in McMaster University’s Orthopaedic Surgery Residency Program

1.4.1 Poor Psychometric Properties of Assessment Tools Currently Used for Evaluating Students on the Upper Levels of Miller’s Pyramid

The psychometric properties of the principal summative mode of assessment used in McMaster University’s orthopaedic surgery residency program, the ITER, have been called into question. For instance, the ITER has shown poor reliability in some studies. Feldman et al. (2004) examined 277 ITERs given to 50 surgical residents. The ITERs assessed whether residents were superior, satisfactory, borderline, or unsatisfactory for 16 attributes; Feldman et al.’s (2004) study assessed responses to one attribute, technical skills. It was found that most students were rated as superior (37%) or satisfactory (61%) on technical skills (Feldman et al., 2004). Only 1% of students were rated as borderline and none were unsatisfactory (Feldman et al., 2004). There is a variety of factors that influence the unwillingness of faculty members to fail residents or grade them poorly. Cleland et al. (2008) conducted focus groups with faculty members in two medical schools in the United Kingdom to determine what these factors may be. They found that some of the faculty members that they sampled had a tendency to underreport underperformance in students that they liked (Cleland et al., 2008). The faculty members also reported a fear of negative consequences of reporting underperforming students, such as stigmatization of those students, the view that failure is destructive to students, and the possibility of a formal appeal (Cleland et al., 2008). The faculty members also identified issues of self-efficacy, such as self-blame for students' failures, lack of confidence that it is the students who are failing and not the curriculum that has failed to adequately prepare the underperforming students, and faculty members' feelings that they cannot give negative feedback effectively or lack adequate documentation to do so (Cleland et al., 2008). The inability of faculty members to fail students decreases the reliability of the results of the ITER since the ability of an assessment tool to discriminate between students is a component of reliability (Streiner and Norman, 2008). The reliability of the ITER was formally assessed in a study of McMaster University's radiology residency program, using scores from 111 ITERS for 14 residents across three rotations (Finlay et al., 2006). The internal consistency of the ITER was found to be 0.91 in this study (Finlay et al., 2006). When items are so highly correlated, it suggests that many items are likely redundant, and it is also likely that faculty are scoring residents based on a global impression rather than discriminating between the particular items on the assessment form (Streiner and Norman, 2008; Finlay et al., 2006). The correlation across rotations was found to be very low, 0.36 (Finlay et al., 2006). This correlation is quite low for what is considered necessary for a high-stakes assessment, which should aim to have a reliability of 0.70 at the very least (Streiner and Norman, 2008). Even when the scores were averaged across the three rotations, as a means to measure competence in general, the reliability remained low, at 0.62, for a high-stakes assessment (Finlay et al., 2006). Furthermore, the completion of ITERs is often not done immediately; they tend to be completed weeks after students end a rotation (Turnbull and Barneveld, 2002).

Clinical encounter cards have been shown to suffer some of these same pitfalls. Richards et al. (2007) examined 7,308 clinical encounter cards for 201 students who completed their surgical clerkship rotation at the University of Texas Health Sciences Centre. Of those 7,308 clinical encounter cards, only 2 were graded as unsatisfactory and 8 as below average (Richards et al., 2007). The remainder were rated average (572), above average (1,699), outstanding (3,211), and intern level (1,807). Nine were missing (Richards et al., 2007). The internal consistency was found to be high at 0.914 in one study evaluating 9,146 clinical encounters that were scored using encounter cards for 50 clerks rotating through internal medicine, surgery, obstetrics and gynaecology, and paediatrics (Al-Jarallah et al., 2005). This high internal consistency suggests that raters act on a global impression when completing these evaluations or that items on the evaluation are redundant. The study by Al-Jarallah et al. also examined the validity of scores generated on clinical encounter cards. They compared scores on clinical encounter cards to global ratings by faculty based on an overall impression of the clerks' competence. The assumption being tested was that clinical encounter cards actually do measure competence, and the assumption would be supported by the finding that scores on clinical encounter cards are similar to those of another measurement of competence, i.e., faculty rankings. Al-Jarallah et al. (2005) found that Spearman's rank-order correlation coefficient was low, at 0.337, which suggested that the ITER scores were not very consistent with the faculty rankings. However, this was not adjusted for the unreliability of the ITER or the faculty rankings, nor was the reliability of each of these assessments calculated.

In light of the psychometric weaknesses of existing measures used to evaluate the higher levels of Miller's pyramid, and in particular the requirement by the RCPSC to use the ITER for the summative assessment of action in spite of its poor psychometric properties, there was a need in the orthopaedic surgery residency program at McMaster University to introduce an adjunct tool with better psychometric properties to assess the upper levels of Miller's pyramid. The measurement principles of the OSCE have shown that it may potentially be used to fill this gap in assessment.

1.4.2 Measurement Principles of the OSCE

The OSCE adheres to several important principles of measurement. According to Van der Vleuten's (1996) utility theory, in order for an assessment to be useful, it should produce reliable and valid scores. As mentioned in Section 1.3, reliability refers to the discriminatory power and consistency of a measurement, and validity refers to evidence supporting the intended interpretation of the scores. A measurement tool should also account for content specificity. Context specificity refers to the finding that the ability to demonstrate competence in one clinical problem has not shown to be predictive of that ability for other clinical problems (Eva, 2003; Norman et al., 1985). Context specificity can be accounted for through sampling multiple problems (Eva, 2003). It should be possible to standardize an assessment tool so that all trainees are assessed based on the same clinical problems if their scores are to be ranked. As discussed in Section 3.1, it must be feasible to implement an assessment tool in the context in which it is intended to be used. An assessment tool must be appropriate for its use in either formative assessment or summative assessment.

OSCEs have consistently been shown to adhere to these measurement principles. For instance, they can have high reliability for many populations, such as medical students and residents at various levels of training and residents in various fields, typically with reliability estimates over 0.60 (Auewarakul et al., 2005; Chipman et al., 2011; Hatala et al., 2011; Yudkowsky et al., 2004).

It is important to note that the reliability of a measurement tool can change when it is applied to a different population (Streiner and Norman, 2008). For example, a study by Petrusa et al. (1990) reported a generalizability coefficient of 0.26 for an OSCE, whereas a study by Yudkowsky et al. (2004) found a generalizability coefficient of 0.81 for another. Reliability can be affected by any source of variance such as the raters used, the number of stations and the length of the examination, the questions asked at stations, and the group of examinees.

Norman et al. (1985) reported that performance is context dependent: performance on one problem is not predictive of performance on another problem. Van der Vleuten et al. (1990) found that the correlations of the measurement of skills across different clinical problems portrayed by standardized patients ranged from 0.1 to 0.3. Further, Norman et al. (1985) identified that this context specificity could even be demonstrated when an identical clinical problem was portrayed by different standardized patients, who were trained to present the problems in an identical way. In this study, residents and clinical clerks were evaluated on problems portrayed by standardized patients over two half-days separated by a period of two weeks (Norman et al., 1985). Of the ten clinical problems, two sets of the same problems were portrayed by different standardized patients on alternate weeks (Norman et al., 1985). Eva (2003) averaged the correlations between the presentation of the same clinical problem by different standardized patients (reported in Norman et al., 1985), and found that it was 0.28. This finding demonstrates the extent to which small variations in the context can affect the rankings of examinees’ performance. In order to overcome context specificity, it is necessary to present multiple clinical problems so that reliable assessments of students’ performance can be made (Eva, 2003).

OSCEs can help overcome the issue of context specificity because their use of multiple stations allows for the presentation of multiple clinical scenarios.

OSCEs also provide an opportunity to standardize assessment in order that all examinees are scored on their performance on the same tasks, which can be difficult to achieve in a clinical setting. There is much flexibility in designing the examination, allowing programs to work within their resource limitations (Harden, 1988). Curriculum designers can overcome resource limitations by designing OSCEs with fewer stations (provided that this does not significantly affect reliability), or developing stations that do not use standardized patients (and instead having interviews with faculty or trainees at a higher level than those being examined), for example. The number of stations may be increased if a program must accommodate many students, or students can be grouped into circuits that run through the same examination at different times or different locations.

These measurement principles suggest that the OSCE may be a psychometrically sound adjunct to current methods of assessing residents, provided that it may be feasibly implemented into the orthopaedic surgery residency program at McMaster University.

1.4.3 Similarity of the Format of McMaster University's Orthopaedic Surgery Residency Program's OSCEs to the RCPSC Comprehensive Objective Examination

All residents must complete the RCPSC certification examination after their fifth year of residency. They complete the first part of the exam, the Surgical Foundations Examination, after a minimum of two years of residency (Royal College, Format of the Comprehensive Objective Examination in Orthopaedic Surgery, 2013). They complete the second part, the Comprehensive Objective Examination, at the end of their training (Royal College, Format of the Comprehensive Objective Examination, 2013). The Comprehensive Objective Examination is composed of a written component and an OSCE component (Royal College, Format of the Comprehensive Objective Examination, 2013). The written component consists of 115 MCQs and 40-60 SAQs (Royal College, Format of the Comprehensive Objective Examination, 2013). The OSCE component consists of approximately 11 stations in the form of structured orals, critical appraisal, telephone consultations, and visual recognition (i.e., SAQs about radiographs or procedures) (Royal College, Format of the Comprehensive Objective Examination, 2013). Prior to introducing the OSCE as an assessment tool into the orthopaedic surgery residency program at McMaster University, faculty observed that residents tended to be concerned about this examination, in particular about the OSCE component. Residents had not been tested with the OSCE in their residency training, were unfamiliar with this testing format in the context of orthopaedic surgery, and were hesitant about what would be expected of them in the Comprehensive Objective Examination OSCE. In order to help senior residents in their preparation for the Comprehensive Objective Examination OSCE, the OSCE was implemented as an evaluative tool in the orthopaedic surgery residency program. Junior residents were tested as well because faculty felt that the OSCE could be an informative tool to assess residents at this stage of learning. Including junior residents in the assessment allows for continuous assessment throughout the residency program, allowing faculty to measure residents' progression through the program in terms of their ability to demonstrate their skills and knowledge.

Since the Comprehensive Objective Examination's OSCE component includes knowledge-testing and performance-testing components, McMaster University's orthopaedic surgery residency program designed a hybrid OSCE to provide residents with a similar examination format by including both knowledge- (MCQs, SAQs) and performance-testing (structured oral interviews) stations. The decision to include a test similar in format to the RCPSC Comprehensive Objective Examination OSCE is supported by evidence that suggests that having experience with the format of a test can improve subsequent pass rates. For example, McMaster University's Undergraduate MD Program did not formally test students with MCQs prior to 1990 (Norman et al., 2010). The failure rates of McMaster University's graduating medical class for the Medical Council of Canada Qualifying Examination Part I (MCCQE I), were consistently below the national average since the second class had graduated and had exceeded four times the national average by 1989, at 19% (Norman et al., 2010). Norman et al. provide data starting from 1981 that suggest that students' average scores had been steadily declining until 1989. In 1990, the program attempted to address this issue by introducing a 180-item multiple choice test, the same testing format as the MCCQE I, which was called a progress test and administered to students three times yearly (Norman et al., 2010). When the progress test was first introduced in 1990, the average score on the LMCC increased significantly and continued to rise steadily (Norman et al., 2010). This suggested that providing students experience with a testing format similar to a high-stakes examination can result in improved scores on the high-stakes examination.

1.4.4 Reasoning for Lack of Consideration of Other Methods with Similar Measurement Principles to the OSCE

There are a couple of performance-testing formats that could have been considered that allow for multiple sampling and have demonstrated good reliability, validity, and feasibility in multiple studies. These testing formats may be more feasible than the OSCE in this setting. One feasibility challenge that was experienced with the OSCEs in the orthopaedic surgery residency program at McMaster University was gathering enough faculty assessors because several faculty members were scheduled to work in the operating rooms on the residents' half-days during which the OSCEs took place.

One potential examination that could have been introduced into the program is the mini-CEX. This consists of a short (10-20 minute) observation by faculty of a resident-patient interaction. The resident is scored in a variety of domains of competence, such as the seven CanMeds roles, and feedback is provided immediately after the encounter. Assessments may be for a specific encounter chosen by the faculty evaluator, or may be spontaneous and unscheduled. Sampling multiple mini-CEX cases allows for overcoming context specificity. Studies have assessed the validity of the use of the mini-CEX in various settings. For example, scores have shown to increase with level of training (Kogan et al., 2003). The mini-CEX has also shown to strongly correlate with other assessments such as multiple choice examinations (Durning et al., 2002). It has been shown in multiple studies that the mini-CEX may provide a reliable assessment of trainees' performance in a clinical setting (Pelgrim et al., 2011). However, it requires approximately ten assessments to achieve a reliability greater than 0.70, which makes it virtually impossible to observe residents on the same clinical scenarios in order to compare residents to each other (Cook et al., 2010; de Lima et al., 2011). Therefore, the OSCE is preferred to the mini-CEX since it has also demonstrated good psychometric properties, but can also compare residents on the same scenarios and can be designed to be similar in format to the RCPSC Comprehensive Objective Examination OSCE.

Patient management problems (PMPs) may be another alternative. This examination is computer-based, and examinees work through a simulated clinical encounter. The encounter may provide a history, report the results of a physical examination, and provide baseline lab values, and the examinees describe the tests and management options they recommend; later problems may introduce related complications (Norcini et al., 1985). It is possible to overcome context specificity through sampling multiple PMPs, and the

PMP has shown to correlate well with knowledge tests in multiple studies (Reddy & Vijayakumar, 2000, Van der Vleuten & Newble, 1995). This testing format is similar to some of the stations in the RCPSC Comprehensive Objective Examination's OSCE component. However, this would not allow residents to practice the other station types, nor would they gain experience with the examination format. The OSCE is preferred and can incorporate PMPs through structured oral interviews.

1.4.5 Feasibility

OSCEs can be resource-intensive, which can present a challenge for programs with significant resource constraints. OSCEs can cost 5 to 70 CAD per student (Walsh et al., 2009). The cost may vary depending on a number of factors, such as the amount of time and resources placed into examination development, the length of the examination, the number of examinees, the number of evaluators, the number of stations, and whether or not standardized patients are used. It is also time-consuming to prepare an OSCE. Cusimano et al. (1994) reported on the time needed to plan and conduct a six-station OSCE for 40 students. Sixty-six hours were spent on station development alone for each OSCE (Cusimano et al., 1994). The exam may run for several hours, depending on how many stations are included and how many students are evaluated. In the six-station OSCE Cusimano et al. (1994) evaluated, 18 surgeon examiners were required for 3.5 hours each (63 surgeon hours total). Each OSCE also required a total of 42 hours of support staff time (Cusimano et al., 1994). Support staff is needed for preparing materials and setting up stations, keeping track of time, helping to rotate students through stations, and entering scores. If standardized patients are involved, training and utilizing them requires even more time: a total of 82.5 hours per OSCE session were reported by Cusimano et al. (1994).

The introduction of the OSCE into the orthopaedic surgery residency program at McMaster University was met with several logistical challenges. One challenge was that OSCEs often had to be scheduled during resident half-days, when several faculty members were in operating rooms, leaving them unavailable to act as evaluators. Adding to the challenge of finding an adequate number of evaluators was the intent to design the OSCEs to be at least eight, and up to ten, stations long. Further, it was decided that the OSCEs would be administered three to four times per year. This issue was overcome by designing half of the OSCE stations to be knowledge-testing (MCQ, SAQ) stations that did not require an evaluator to be present, reducing the amount of evaluators to recruit at each OSCE to half of the original expected amount. Keeping an up to date question bank was a challenge and is an ongoing challenge, since treatment algorithms tend to change over time. Having a large enough question bank was also challenging, since residents expect to be tested using different questions as they complete each OSCE. A final challenge was designing questions that were appropriate to the residents’ levels of training at each OSCE.

1.5 Misconceptions about the OSCE

It is necessary to clarify a few misconceptions that, at times, arise in studies that discuss OSCEs. When OSCEs were originally introduced by Harden et al. in 1975, it was suggested that an advantage of the exam was that OSCE developers could take many measures aimed at making the test more objective (Harden, 1988). Some steps suggested were the use of checklists, training evaluators, and the standardization of criteria between evaluators (Harden, 1988). These measures were expected to minimize the effects of biases from raters, making the measurement of performance strictly that: a measure of the examinee's performance, without subjective influences from the examiners (Harden, 1988). However, making an examination more objective (which is called objectification) does not necessarily improve a test's psychometric properties. An example that demonstrates this is the use of checklists over global ratings to evaluate students' performance at stations. A review of the literature summarized studies reporting the reliabilities of checklists and global ratings, and found the reliabilities of the measures to be comparable (van der Vleuten et al., 1991). This illustrates that objectified methods of assessment are not necessarily more reliable than subjective ones (van der Vleuten et al., 1991). Further, scores on objectified methods of assessment have not necessarily shown to be more valid for their intended interpretations. A review by Van Luijk and Van der Vleuten (1991) found that the correlation between global ratings and checklist scores in an OSCE administered to medical students in the Faculty of Medicine in Maastricht, The Netherlands, were 0.81. This result suggests that not very much additional information about an examinee's performance would be gained through the use of the relatively less subjective method, checklists, than with global ratings (Norman et al., 1991).

Another misconception about the examination is the ability of the OSCE to be used as a means of generating a profile about a candidate's competence in many different specific areas of competency, or as a means of evaluating whether the curriculum is teaching important concepts adequately (Walsh et al., 2009). It has been suggested that all stations assessing one particular domain of competence (e.g., communication) can be used to generate the examinee's score in that particular domain. If a resident scores poorly on stations covering one domain of competence, this implies that they are week in that domain; if all residents score poorly on those stations, this implies that the curriculum is not adequately preparing them in that domain. However, this is not the case. Stations are often the largest source of error in the OSCE, and it follows that the reliability of an OSCE increases with the number of stations in it. If individual competencies are only tested with a few stations each, then there are not enough stations to construct a reliable conclusion about examinees' performance on individual competencies. The OSCE can only provide an overall impression of examinees' performance.

1.6 Methods of Scoring

In an OSCE, examinees are generally scored with checklists or global ratings. There are no differences between the two scoring methods in terms of their reliability, and their scores have been demonstrated to correlate highly (van Luijk and van der Vleuten, 1991; van der Vleuten et al, 1991). However, it does not appear that checklists are able to differentiate between increasing levels of expertise (Hodges et al, 1999). Participants in a study by Hodges et al completed two OSCE stations in which they interviewed two standardized patients (1999). They were rated with the use of a 22-item binary checklist and a global rating scale encompassing five domains (knowledge and skills, empathy, coherence, verbal expression, and nonverbal expression) (Hodges et al, 1999). Examinees were in one of three groups: clinical clerks, family practice residents, and family physicians (Hodges et al, 1999). When measured by global ratings, physicians scored better than residents or clerks; however, the scores of the binary checklists demonstrated a downward trend as level of expertise increased (Hodges et al, 1999). This trend likely occurred because experts manage clinical problems through the gathering of focused information, whereas novices gather a lot of data without as much consideration for context, as if working through a checklist themselves (Hodges et al, 1999). As a result of this study, Hodges et al (1999) did not recommend the use of binary checklists for evaluating groups in which some individuals may have achieved more expertise than others. However, the examinees in this study were assessed only on psychiatry stations and these stations were designed challenge clinical clerks. Thus, the results may not necessarily be generalizable to all OSCE examinations. Further, there may be instances when checklists are preferable to global ratings, such as when evaluators are not experts in what is being tested. For example, Norcini et al. (1990) evaluated an examination by the American Board of Internal Medicine that tested internal medicine residents with 12 questions that necessitated a short essay response. Three non-medical raters, who were trained for 14 hours, evaluated the responses with a checklist, and two fellows in internal medicine provided a global rating on a 9-point scale (Norcini et al., 1990). The correlation between the scores generated by the non-medical raters and those generated by the fellows was 0.87 (Norcini et al., 1990). It appears that when raters use checklists to evaluate content that is outside of their scope of expertise, they are able to predict well the scores that would be generated by content experts.

1.7 Research Objectives

As mentioned, there are many challenges to feasibly implementing a resource-intensive examination such as the OSCE into orthopaedic surgery residency program. The use of the OSCE has not been studied extensively in the setting of orthopaedic surgery residency training, and there is little evidence to support that this examination may be introduced in a feasible manner, and OSCE scores have not demonstrated to be reliable and valid in other orthopaedic surgery residency programs.

A review of the literature produces few articles on OSCEs in the setting of orthopaedic surgery residency programs. Shaheen et al. (1991) provide a preliminary discussion on the development of an OSCE prior to its implementation in the orthopaedic surgery residency program at King Saud University in Saudi Arabia. Since it is only a preliminary discussion, there is no information on how successfully the OSCE was implemented into the program, such as information regarding feasibility issues that did or did not exist when the OSCE was implemented and how they were overcome, and whether the scores on the OSCE were reliable and valid for their intended interpretation.

In the Ohio State University orthopaedic surgery residency program, musculoskeletal physical examination skills are assessed with OSCE examinations (Beran et al., 2012; Griesser et al., 2012). Published studies discussing these OSCEs do not address logistical challenges. They mention that subspecialist evaluators reviewed videos of stations that covered their specialty area after the OSCE took place, but do not discuss whether this proved to be feasible or presented any of its own unique challenges such as a significant delay in providing scores to residents (Beran et al., 2012; Griesser et al., 2012). These studies also do not indicate whether the OSCE provided reliable results in this population (Beran et al., 2012; Griesser et al., 2012). The OSCEs at The Ohio State University only included musculoskeletal physical examinations, whereas the RCPSC Comprehensive Objective Examination OSCE component includes a variety of other stations that residents do not have the opportunity to practice in this setting. McMaster University's orthopaedic surgery residency OSCEs are unique in their attempt to produce a similar examination format to the OSCE component of the Comprehensive Objective Examination. This allows residents to gain experience with this examination format and thus provides them the potential to score higher than they might have had they not had any familiarity with the examination format.

There is little evidence to understand the implications of an OSCE, and in particular one that is a hybrid between knowledge- and performance-testing stations, in the setting of an orthopaedic surgery residency program. Although OSCEs are not well-studied in this setting, they are widely used in Canadian orthopaedic surgery residency programs and are seen as important to the assessment of residents' performance by program directors and residents. The results of a survey administered program directors and residents in orthopaedic surgery revealed that 91% of program directors and 67% of residents in Canada who responded to the survey rated OSCEs as at least somewhat important on a 7-point Likert scale (1 = not applicable, 2 = extremely unimportant, 3 = very unimportant, 4 = somewhat unimportant, 5 = somewhat important, 6 = very important, and 7 = extremely important) (Evanview, 2013). Further, 58% of program directors in Canada who responded to the survey rated the OSCE as either very important or extremely important (Evanview, 2013).

This study attempts to address the gap in the literature that exists in the evaluation of the utility of OSCEs in the setting of orthopaedic surgery residency programs. Utility theory was used to assess the OSCEs in McMaster University's orthopaedic surgery residency program (Van der Vleuten, 1996). As described in Section 3.1, utility theory requires evaluating an assessment tool in five domains: reliability, validity, feasibility, acceptability, and educational impact (Van der Vleuten, 1996).

The analyses included four OSCEs that took place between October 2010 and September 2011. One is a junior OSCE which took place in June of 2011, and the remaining three are senior OSCEs which took place in October of 2010, and March and September of 2011. Although PGY 3 residents are senior residents, they were required to take part in the junior OSCE rather than the senior OSCEs; however, they were permitted to participate in senior OSCEs if they desired to do so.

1.7.1 Primary Objective

The primary aim of this study was to inform the orthopaedic surgery residency program at McMaster University of the reliability of the OSCEs that took place in the 2010-2011 and 2011-2012 academic years.

1.7.2 Secondary Objectives

Construct validation studies were included as secondary objectives. OSCEs were evaluated to determine whether residents' scores ranked as expected (i.e., whether residents of higher PGYs scored higher than residents of lower PGYs on the same OSCEs).

Due to limitations in the availability of faculty evaluators, approximately half of the OSCE stations were knowledge-testing stations (MCQs, SAQs, radiographs to diagnose).

These station formats were used to maximize the representation of the format to the RCPSC certification examination, thus helping students better prepare. Another secondary objective of this study was to determine if residents' scores on the OSCEs' knowledge-testing stations correlated with performance-testing stations. In addition, reliability of both station formats was tested.

Another secondary goal was to determine the residents' acceptability of the OSCE, which was estimated with exit surveys. Residents' comments on these surveys also informed educational impact, because they could comment on how implementation of the OSCE changed such factors as study habits.

Chapter 2: Methods

2.1 McMaster University's Orthopaedic Surgery Residency Program OSCEs

The orthopaedic surgery residency program began to use OSCEs as a method of formative evaluation in October of 2010. This first OSCE was followed by another senior OSCE in March 2011, assessing the same cohort of students, and a junior OSCE in June 2011. In September 2011, a new cohort of PGY 4 residents and the previous cohort of PGY 4 residents, now PGY 5 residents, participated in a senior OSCE. Attendance at the relevant OSCEs was mandatory for all residents.

The OSCE in October 2010 was eight stations long. The OSCEs in March 2011 and September 2011 (senior OSCEs) were ten stations long. The OSCE in June 2011 (junior OSCE) was nine stations long. Residents spent nine minutes at each station, with one minute between stations to read the stems. Each OSCE was conducted with two circuits to accommodate the number of residents, one circuit immediately following the other to prevent sharing of information between residents participating in different circuits. Exit surveys were completed by residents after each OSCE.

2.2 OSCE Development

Faculty members of the orthopaedic surgery residency program were responsible for question development. Faculty members contributed questions relevant to their individual subspecialties to the program director. The program director and associate program director reviewed the proposed questions and accepted them based on their appropriateness to the residents' level of training. Since 2012, the orthopaedic surgery residency program has designated an OSCE director who is expected to collect OSCE questions and verify their appropriateness to residents' level of training.

Questions for the senior OSCE were designed to be challenging to both PGY 4 and PGY 5 residents, and were designed to be the same difficulty as what would be expected on the RCPSC Comprehensive Objective Examination OSCE component. Junior OSCE stations were designed to challenge PGY 1 and PGY 2 residents; PGY 3 residents were expected to find these stations less challenging.

2.3 Blueprinting

The OSCEs were developed with the orthopaedic surgery postgraduate curriculum in mind, testing important concepts that residents were expected to learn. Blueprinting is important in developing an examination for two reasons. Firstly, it is fair: it results in residents being tested on concepts that have been stressed by faculty, and more important concepts are given more coverage. It also demonstrates what the important concepts in the curriculum are and therefore encourages the study of those concepts. Students study what they think will be assessed and what is assessed in turn identifies what is important to study (Van der Vleuten, 1996).

2.4 Station Types

The stations developed for each OSCE were of two formats: knowledge-testing stations and performance-testing stations. As mentioned, the knowledge-testing stations were included for two reasons: 1) limited availability of faculty examiners necessitated the development of examiner-free stations; and 2) this was reflective of the RCPSC certification examination, for which these OSCEs are intended to prepare the residents. The RCPSC Comprehensive Objective Examination OSCE contains both knowledge- and performance-testing stations.

Appendix 3 provides examples of questions from each type of station.

2.4.1 Knowledge-Testing Stations

MCQs

Multiple choice questionnaires were developed for this examination, and these tested a variety of concepts at a time. At the October 2010 and March 2011 OSCEs, there were 25 and 23 questions, respectively. However, faculty and residents found this to be too long for a ten-minute station and therefore reduced the length of the multiple-choice questions to ten per MCQ station for subsequent OSCEs. Typically, only one MCQ station was included in each OSCE.

SAQs

Short answer questions were typically short, open-ended questions that could be answered with only a few words or a few lines. Unlike MCQs there were no options present. Typically, short answer questions presented cases from a particular domain of orthopaedics (e.g., paediatrics), with follow-up questions based on the cases. The use of SAQs avoids the potential problem of cueing, in which the correct answer is recognized in the options, but potentially may not have been provided by the examinee if no options were provided (Epstein, 2007).

Spot Diagnosis

At these stations, residents were presented with 11 radiographs and were required to provide diagnoses.

2.4.2 Performance-Testing Stations

Oral Examinations

At these stations, faculty interviewed residents. They provided residents with information on a case, and began a conversation exploring diagnostic and case management issues. Throughout this conversation, faculty provided information which had the potential to affect case management, and observed how each resident responded to the new information. The faculty member that interviewed the residents at each station served as the evaluator.

Technical Stations

Residents were presented with a setup that required them to simulate performing a procedure. They were provided with surgical tools, and models were set up on which they would operate. A faculty member was present, either in the room or behind a one-way window, observing the operation. This faculty member served as the evaluator.

MMI Stations

One station of one OSCE (September 2011) was designed in this format. This station tested residents on the CanMEDS competency professionalism. A hypothetical scenario was provided, and the resident discussed how they would resolve the conflict that was presented.

2.5 Scoring

2.5.1 Knowledge-Testing Stations

For MCQs, the proportion of correct answers out of total questions was converted into a proportion out of 100. The proportion of correct answers in each SAQ station was also multiplied into 100. For spot diagnosis stations, the proportion of correct diagnoses was multiplied into100.

2.5.2 Performance-Testing Stations

Given the assumption that residents develop more expertise as they progress through residency, global ratings were selected over binary checklists to score residents in performance-testing stations. Global ratings were seen as the more efficient to design of the two scoring methods, since the creation of detailed checklists for each oral OSCE station would potentially be a time-consuming task. Given that the two methods correlate highly, the scoring method that is more efficient to design is preferable.

Scoring changed over time, reflecting changes in expectations of the faculty. Appendix 4 outlines the scoring of performance-testing stations in each OSCE. For all performance-testing stations, each resident's total score was divided by the maximum possible score for each station and multiplied into 100 to achieve the station score.

2.5.3 Total OSCE Score

After all of the individual station scores were converted into a score out of 100, they were added to produce the total OSCE score. Residents failed the OSCE if their score was less than 60%. The cutoff of 60% was chosen by the program director and assistant program director because it typically resulted in one or two examinees failing, identifying those who were below their peers. Although the OSCE was designed as a formative assessment, residents were given feedback that they failed the OSCE. This was done because the only form of summative assessment used to assess residents, the ITER, has not shown the ability to discriminate between residents who are performing well and those who are performing poorly compared to their peers (Feldman et al., 2004; Finlay et al., 2006).

2.5.4 Rater Selection and Training

Evaluators were faculty members who volunteered to participate in evaluating residents for the OSCEs. In the junior OSCE, senior residents (with the exception of PGY 3 residents) also volunteered as evaluators. Faculty members typically were supportive of the use of the OSCE to evaluate residents and participated unless they had a commitment in an operating room. Half of the raters were assigned to a station in which they were content experts, and the other half were not. Although it was ideal to have content experts evaluate stations that tested their domain of expertise, this was not consistently done so that residents were unable to predict the content of the station upon seeing the examiner.

Raters were not trained to evaluate residents. However, if they were not experts on the content of their station (i.e., if a particular rater was a trauma surgeon but was evaluating a paediatrics station), they discussed the station with someone who was an expert in the content. Additionally, they were given a list of major discriminators that were expected to be mentioned by the resident at that particular station.

2.5.5 Feedback

Residents were provided with their scores on the OSCE and each individual station. Select stations would be discussed at rounds following the OSCE.

2.6 Exit Surveys

The exit surveys that were administered after each OSCE contained the following questions:

1. Do you think you were able to present an accurate portrayal of your ability during the OSCE?
2. Compared to a traditional MCQ/SA test, do you think the OSCE causes more or less anxiety?
3. Was the process more or less stressful than you anticipated?
4. Were the questions given before each station adequate preparation?
5. In general, do you think the stations covered appropriate topics?
6. In general, how difficult were the stations?
7. In general, was the time available for the stations appropriate?

The following 7-point Likert scale, including an additional option for not applicable (N/A), was used:

1	2	3	4	5	6	7	8
Definitely not		Not really		Somewhat		Definitely	N/A

The anchors on the scale (definitely not, not really, somewhat, and definitely) were appropriate for questions 1, 4, 5, and 7, but not for questions 2, 3, and 6. Although it appears that the residents generally understood, for example, that for question two, a higher number on the scale referred to increased anxiety (given the consistency in the scores and the notes that some residents left on the scale indicating how they interpreted it), the questions for which the anchors were inappropriate were analyzed separately from the other questions. This was done because of the potential for the residents' interpretations of the anchors to have varied for the questions for which the anchors were not appropriate. For the analysis, a higher score on question 2 was interpreted to mean more anxiety, a higher score on question 3 was interpreted to mean more stress, and a higher score on question 6 was interpreted to represent a greater degree of difficulty. Residents completed the exit surveys anonymously.

2.7 Missing Data

If a station score was missing for a resident, the resident's mean score for the remaining available stations replaced the missing station score. Some stations that were scored with multiple components, such as oral stations in which residents were rated in multiple categories, had missing scores for one of the categories. In the example of an oral station score, a missing score in one category was replaced by the mean score of the remaining categories. If a rating for a question on the exit survey was missing, that question was ignored and the entry was left blank. Any analyses of the exit surveys ignored the missing data (i.e. mean score for a question was calculated only for the available data points).

2.8 Plans of Analysis and Hypotheses

The analysis was done following the 5 domains of utility theory outlined by Van der Vleuten in 1996: reliability, validity, acceptability, feasibility, and educational impact. Data are available to inform reliability, validity and acceptability. Measures were undertaken to make the examination more feasible, and these will be discussed. There are limited data available to determine educational impact at this time; there were a number of comments on exit surveys that informed this domain of utility theory.

2.8.1 Reliability

The importance of measuring reliability is determining the extent to which both random and systematic error exists in a measurement tool when it is applied to a particular population (Streiner and Norman, 2008). Reliability was assessed under generalizability (G) theory. G_STRING IV software was used to generate the G coefficient (the intraclass correlation coefficient (ICC)), based on calculations of variance components by urGENOVA software. The reliability of each OSCE was assessed, as well as the reliability of the knowledge-testing and performance-testing stations of each OSCE.

The following steps were followed in the analysis (Streiner and Norman, 2008):

Step 1: Identifying All Potential Sources of Variance

In this design, variance is contributed by residents, stations, raters, and the year of residency.

The facet of differentiation is the object of measurement. In this case, residents were the facet of differentiation. The facet of stratification categorizes the object of measurement, and year of residency was the stratification facet. The facets of generalization are the sources of error variance. When generating variance components and G coefficients for the OSCEs overall and the performance-testing stations, the variance contributed by rater and station and their interaction could not be separated from each other, since there was only one rater per station. The facet of generalization, then, was the combination of rater and station and their interaction. Since there were no raters in the knowledge-testing stations, station was the facet of generalization. Residents were nested (:) in the year of residency. All remaining facets and interactions between facets were crossed (×).

Step 2: Calculating the variance components:

urGENOVA generated the variance components. Each facet and the interactions of the facets constituted the variance components: Resident:Year, Year, Station (station and rater and their interaction) , Year×Station (station and rater and their interaction), and Resident:Year×Station (station and rater and their interaction).

Step 3: Calculating the G Coefficient

G_STRING IV was used to calculate the G coefficients from the variance components. The relative error coefficient was chosen over the absolute error term because the scores are to be interpreted relative to each other rather than as the absolute values (Bloch and Norman, 2011). The formula for calculating the relative error coefficient, into which the variance components were substituted was:

$$G = \frac{\sigma(\tau)^2}{\sigma(\tau)^2 + \sigma(\delta)^2}$$

(Bloch and Norman, 2011). This equation generates the ICC, with τ representing the variance associated with the facet of differentiation and δ representing the error variance (Bloch and Norman, 2011). The main effects were not included in the relative error coefficient (Bloch and Norman, 2011). Interactions with the facet of stratification were also not included (Bloch and Norman, 2011). Therefore the variance component contributing to the error was P:YxS, generalized over *n* total stations. The G coefficient was therefore calculated as follows:

$$G = \frac{\sigma(P)^2}{\sigma(P)^2 + \frac{\sigma(P\!:Y \times S)^2}{n}}$$

(Bloch and Norman, 2011).

The desired level of the generalizability coefficient that was chosen was at least 0.70 (Streiner and Norman, 2008). This level of generalizability is expected in a high-stakes evaluation. The only form of high-stakes assessment in the program is the FITER, and for each rotation the ITER; as discussed, this can be an unreliable testing format (Feldman et al., 2004; Finlay et al., 2006). Since one of the reasons for including the OSCE in the assessment of residents was to have a tool with superior psychometric properties to existing high-stakes evaluations of the higher levels of Miller's pyramid, the reliability of the OSCE is therefore expected to be high. It is not necessary to have a higher G coefficient because this OSCE is not a high stakes examination (i.e. its results do not immediately result in any decision-making process, unlike the RCPSC examination whose results factor into determining whether a resident will receive certification).

Step 4: Decision (D) Study

A Decision (D) study was only run when the G coefficient did not reach 0.700. In this case, the number of stations was increased in G_STRING IV until the G coefficient reached the desired level of reliability. This would determine the number of stations that the OSCE would need to achieve a reliability coefficient of 0.700.

Hypotheses

1. Each OSCE will demonstrate to be reliable for use in this population. Overall, G coefficients will be approximately 0.60 or greater, as predicted by the literature (Chipman et al., 2011; Hatala et al., 2011; Yudkowsky et al., 2004).
2. The knowledge-testing and performance-testing stations will be reliable for use in this population. However, because the generalizability of the OSCE increases with

the number of stations, the generalizability coefficient may not reach the desired level (0.700) because there are only four to six stations of each kind in each OSCE. A D study will have to be conducted to determine the number of stations that would be needed to achieve the desired level of reliability.

2.8.2 Validation Tests

It is important to conduct validation tests to determine the extent to which scores on an assessment tool are measuring that which they are intended to measure in a particular population (Streiner and Norman, 2008). The following describes the validation tests that were conducted:

Construct Validation Test 1:
If the OSCE is measuring residents' performance, those who are expected to perform better should score significantly higher than those who are not expected to perform as well. That is, residents who are in a higher PGY should score higher on the OSCE than residents who are in a lower PGY. This was tested with a two-by-two repeated measures ANOVA in order to gauge the influence of station type on scores as well as resident year on scores. The within subjects factor was station type and the between-subjects factor was the total score on knowledge-testing and performance-testing stations.

Construct Validation Test 2:

The knowledge-testing stations were included instead of performance-testing stations because of limitations in the availability of faculty examiners. This was done under the assumption that the results of knowledge-testing stations would be predictive of the results of performance-testing stations. This assumption was tested.

For each resident, their total score on all performance-testing stations in all of the OSCEs was calculated and converted into a score out of 100. The same was done for the performance-testing stations. The two sets of scores were correlated with an interclass correlation efficient (Pearson's r).

Reliability places an upper limit on validity (Streiner and Norman, 2008). Therefore, the correlation coefficient that is adjusted for the unreliability of the two measures was also determined. For the senior OSCE, the average G coefficients of each station type were calculated for use in this analysis. For the junior OSCE, the G coefficients that were generated in the generalizability analysis were used. The G coefficients and unadjusted Pearson's r were substituted into the following equation:

$$r'_{KP} = \frac{r_{KP}}{\sqrt{r_{KK} r_{PP}}}$$

(Norman and Streiner, 2000). In this equation, $r'_{KP'}$ represents the adjusted correlation, $r_{KP'}$ represents the unadjusted correlation, r_{KK} represents the G coefficient of the knowledge-testing stations, and r_{PP} represents the G coefficient of the performance-testing stations.

The coefficient of determination (R^2) was calculated by squaring r in order to determine the extent to which differences in knowledge accounted for variations in residents' performance. The R^2 value, adjusted for the unreliability of each station type, was determined by squaring the disattenuated Pearson's r.

Hypothesis

There will be a moderate (~0.600) correlation between performance-testing stations and knowledge-testing stations, as predicted by the literature (Matsell et al., 1991).

2.8.3 Acceptability and Educational Impact

The means and standard deviations of the response options of the exit survey were calculated to measure acceptability. Residents were encouraged to submit comments on the exit survey and did so; these comments informed educational impact.

Hypothesis

The OSCE will show to be acceptable to residents on the domains assessed in the exit survey. Previous literature has shown the OSCE to be acceptable to residents (Duerson et al., 2000).

2.9 Research Ethics Board Approval

This study was deemed to be quality assurance by the Research Ethics Board (REB) at McMaster University because it was conducted as part of a retrospective program evaluation. As such, it did not require a formal REB approval.

Chapter 3: Results

Tables and figures are provided in Appendix 2.

3.1 Summary Statistics and Comparisons of Means

Figures 2-5 represent the distribution of scores obtained in each OSCE, converted to a score out of 100. The distribution of scores was plotted in order to demonstrate the extent to which the OSCE does or does not share the shortcoming of other methods such as ITERs and clinical encounter cards, of grouping all residents at the upper end of the possible scores; if it did share this shortcoming, the results should be consistently skewed left. Knowledge- (evaluator free) and performance-testing stations were separated. The October 2010 OSCE curve appears normally distributed for performance-testing stations, and knowledge-testing station scores are skewed left. The scores for the March 2011 and June 2011 OSCEs are normally distributed for both station types. For the September 2011 OSCE, the knowledge-testing stations were normally distributed, while the performance-testing stations were skewed left. In all OSCEs, a greater frequency of higher scores (≥80.00%) were in performance-testing stations than performance-testing stations, and a greater frequency of lower scores (≤50.00%) were in knowledge-testing stations.

Table 1 provides the summary statistics for the October 2010 OSCE. In this OSCE, there were 11 PGY 4 residents and 6 PGY 5 residents, totalling 17 residents who took the examination. This OSCE consisted of eight stations, which covered the following topics: general orthopaedics, arthroplasty, fractures, spine trauma, paediatrics, trauma, tumour, and sawbone (technical skills station). Four tested knowledge (1 MCQ, 3 SAQs) and four tested performance (3 oral, 1 technical). The maximum possible score was 800. The mean score was higher for PGY 5 residents than for PGY 4 residents (mean (SD) = 582.25 (55.18) and 455.12 (70.19), respectively). The total OSCE score ranged from 336.43 to 542.44 for PGY 4 residents and from 493.36 to 655.60 for PGY 5 residents.

Table 2 provides the summary statistics for the March 2011 OSCE. In this OSCE, there were 8 PGY 4 residents and 8 PGY 5 residents, totalling 16 residents who took the examination. This OSCE consisted of ten stations, which covered the following topics: general orthopaedics, foot and ankle, tumours, paediatrics, upper extremity, sports, arthroplasty, and sawbone (technical station). Four tested knowledge (1 MCQ, 3 SAQs) and six tested performance (5 oral, 1 technical). The maximum possible score was 1,000. The mean score was higher for PGY 5 residents than for PGY 4 residents (mean (SD) =712.56 (104.84) and 632.19 (76.08), respectively). The total OSCE score ranged from 510.84 to 725.12 for PGY 4 residents and from 559.48 to 881.74 for PGY 5 residents.

Table 3 provides the summary statistics for the June 2011 OSCE. In this OSCE, there were 5 PGY 1 residents, 8 PGY 2 residents, and 7 PGY 3 residents, totalling 20 residents who took the examination. This OSCE consisted of nine stations, which covered the following topics: general orthopaedics, hip, forearm, leg pain, leg lesion, and sawbone (technical station). Four tested knowledge (1 MCQ, 1 spot diagnosis, 2 SAQs) and five tested performance (4 oral, 1 technical). The maximum possible score was 900. The mean score was higher for PGY 3 residents than for PGY 2 residents, and the mean score for PGY 2 residents was higher than the mean score for PGY 1 residents (mean (SD) = 691.38 (88.04), 655.40 (76.12), and 652.24 (80.78), respectively). The total OSCE score ranged from 554.27 to 754.95 for PGY 1 residents, from 564.97 to 764.06 for PGY 2 residents, and from 521.46 to 777.68 for PGY 3 residents.

Table 4 provides the summary statistics for the September 2011 OSCE. In this OSCE, there were 6 PGY 4 residents and 12 PGY 5 residents, totalling 18 residents who took the examination. This OSCE consisted of ten stations, which covered the following topics: general orthopaedics, upper extremity, hip, tumour, lesion, spine, trauma, paediatrics, and professionalism. Five tested knowledge (1 MCQ, 1 spot diagnosis, 3 SAQs) and five tested performance (4 oral, 1 MMI). The maximum possible score was 1,000. The mean score was higher for PGY 4 residents than for PGY 5 residents (mean (SD) = 804.82 (21.76) and 772.03 (129.26), respectively). The total OSCE score ranged from 780.27 to 835.13 for PGY 4 residents and from 543.38 to 921.01 for PGY 5 residents.

Residents needed to achieve a total score of 60% to pass the OSCEs. Seven of the 11 (64%) PGY 4 and none of the five PGY 5 residents failed the October 2010 OSCE. Three of 8 (38%) PGY 4 and 2 of 8 (25%) PGY 5 residents failed the March 2011 OSCE. None of the 5 PGY 1, none of the 8 PGY 2, and 1 of the 7 (14%) PGY 3 residents failed the

June 2011 OSCE. None of the 6 PGY 4 and 2 of the 12 (17%) of the PGY 5 residents failed the September 2011 OSCE.

3.2 Construct Validation: Comparisons of Means

Figure 6 graphs the mean total scores of all three senior OSCEs, separated by year of residency. The mean scores in October 2010 and March 2011 are higher for PGY 5 residents than for PGY 4 residents, but higher for PGY 4 residents than for PGY 5 residents in the OSCE in the subsequent academic year (September 2011). Figure 7 graphs the mean total scores of the junior OSCE, separated by year of residency. The mean score increases slightly with the year of residency.

The difference in means between PGY 4 and PGY 5 residents in the October 2010 OSCE was 127.13 ($p = 0.001$, $d = -2.01$). The difference in means between PGY 4 and PGY 5 residents in the March 2011 OSCE was 80.37 ($p = 0.103$, $d = -0.88$). The difference in means between the PGY 4 and PGY 5 residents in the September 2011 OSCE was 32.79 ($p = 0.315$, $d = 0.35$). For the junior OSCE in June 2011, a one-way analysis of variance revealed no significant effect of year of residency ($F(2,17) = 0.475$, $p = 0.630$, $\varepsilon^2 = 0.053$).

In order to account for the potential of an interaction between station type and level of training, a two-way ANOVA was conducted to assess the effect of level of training, station type, and the interaction, on residents' scores. In the October 2010 OSCE, the mean scores for knowledge-testing stations (out of 400) were lower for PGY 4 residents than for PGY 5 residents (mean (SD) = 198.54 (43.93) and 248.30 (25.43), respectively). The mean scores for performance-testing stations (out of 400) were also lower for PGY 4 residents than for PGY 5 residents (mean (SD) = 254.33 (44.67) and 318.99 (44.52), respectively). Figure 8 plots the mean scores, separated by station type, for PGY 4 and PGY 5 residents. The main effect of station type was significant (F(1,30) = 34.209, p < 0.001). The interaction between station type and post-graduate year was not significant (F(1,30) = 0.474, p = 0.501). The main effect of the year of residency was significant (F(1,30) = 0.11.027, p = 0.005).

The mean scores for each station type in the March OSCE were transformed to scores out of 100 because of the imbalance between the number of knowledge-testing and performance-testing stations. The mean scores for knowledge-testing stations (out of 100) were lower for PGY 4 residents than for PGY 5 residents (mean (SD) = 53.27 (9.57) and 62.67 (13.54), respectively). The mean scores for performance-testing stations (out of 100) were also lower for PGY 4 residents than for PGY 5 residents (mean (SD) = 68.85 (8.78) and 77.25 (12.24), respectively) and higher than the scores in the knowledge-testing stations in both groups. Figure 9 plots the mean scores, separated by station type, for PGY 4 and PGY 5 residents. The main effect of station type was significant (F(1,28) = 24.420, p < 0.001). The interaction between station type and year of residency was not significant (F(1,28) = 0.063, p = 0.805). The main effect of the year of residency was also not significant (F(1,28) = 3.174, p = 0.096).

The mean scores for each station type in the June 2011 OSCE were also transformed to scores out of 100 because of the imbalance between the number of knowledge-testing and performance-testing stations. The mean scores for knowledge-testing stations (out of 100) were lowest for PGY 1 residents, higher for PGY 2 residents, and highest for PGY 3 residents (mean (SD) = 63.10 (13.01), 65.34 (14.73), and 71.02 (13.19), respectively). The mean scores for performance-testing stations (out of 100) were only slightly higher for PGY 1 residents than for PGY 2 residents, and highest for PGY 3 residents (mean (SD) = 79.97 (5.89), 78.81 (4.54), and 81.46 (8.38), respectively). The mean scores were consistently higher for performance-testing stations than for knowledge-testing stations in all three groups. Figure 10 plots the mean scores, separated by station type, for PGY 4 and PGY 5 residents. The main effect of station type was significant (F(1,34) = 24.420, p < 0.001). The interaction between station type and year of residency was not significant (F(1,34) = 0.063, p = 0.805). The main effect of the year of residency was also not significant (F(1,34) = 3.174, p = 0.096).

In the September 2011 OSCE, the total scores for knowledge-testing stations (out of 500) were higher for PGY 4 residents than for PGY 5 residents (mean (SD) = 379.82 (25.01) and 370.09 (68.69), respectively). The total scores for performance-testing stations (out of 500) were also higher for PGY 4 residents than for PGY 5 residents (mean (SD) = 425.00

(16.52) and 401.94 (71.21), respectively). The mean scores for performance-testing stations were higher than the mean scores for knowledge-testing stations in both groups. Figure 11 plots the mean scores, separated by station type, for PGY 4 and PGY 5 residents. The main effect for station type was significant (F(1,32) 9.947, p = 0.006). The interaction between station type and post-graduate year was not significant (F(1,32) = 0.297, p = 0.593). There was no main effect of the year of residency (F(1,32) = 0.370, p = 0.552).

3.3 Generalizability

Tables 5-16 provide the variance components from each source of variance for each analysis, as well as the percentage of the total variance contributed by each component. All negative variance components are considered to be 0 in the calculation of the G coefficient. The variance components relevant to the calculation were P:Y (τ) and P:Y×S (δ).

Year of residency contributes a significant amount of variance in the October 2010 OSCE (58%) and is larger than any other source of variance (Table 5). When the knowledge-testing stations were examined, it contributed 15.83% of the total variance, and it contributed 48.13% of the total variance in the performance-testing stations (Tables 6-7). The March 2011 OSCE indicates some contribution of variance from year of residency (5.09%) (Table 8). In the generalizability analysis on the knowledge-testing stations, it contributed 4.63% of the total variance and it contributed 3.80% of the total variance in the performance-testing scores (Tables 9-10). The variance components were negative for year of residency in the generalizability analyses of the June and September 2011 OSCEs (Tables 11-16).

Table 17 provides a summary of the G coefficients generated for each OSCE, as well as their knowledge- and performance-testing stations. All OSCEs achieved the desired reliability of ≥0.700. The knowledge-testing stations for the October 2010 OSCE failed to achieve a G coefficient of this size with four stations (G=0.58). A D study found that 7 knowledge-testing stations would be required for the knowledge-testing stations to achieve a G coefficient ≥0.70 on their own (Table 18). The performance-testing station of the October 2010 OSCE achieved a G coefficient of 0.71. No D study was necessary. Both knowledge-testing and performance-testing stations of the March 2011 OSCE had G coefficients of only 0.71 and 0.67, respectively. The 4 knowledge-testing stations would have to be increased to 5, and the 6 performance-testing stations would have to be increased to 9 to each have a G coefficient ≥0.70 (Tables 19-20). The knowledge- and performance-testing stations in the June and September 2011 OSCEs demonstrated G coefficients ≥ 0.70.

3.4 Construct Validation: Correlation of Knowledge- and Performance-Testing Stations

3.4.1 Senior OSCEs

Figure 12 demonstrates the correlation between knowledge- and performance-testing stations, with a line of best fit plotted with the data. The unadjusted correlation between all knowledge-testing and performance-testing stations completed by senior residents was 0.71, and the unadjusted R^2 value was 0.50. The average generalizability of senior OSCE knowledge-testing stations was 0.66 and for senior OSCE performance-testing stations was 0.72. The correlation, adjusted for the unreliability of the two examinations was 1.00. The adjusted R^2 value was 1.00.

3.4.2 Junior OSCEs

Figure 13 demonstrates the correlation between knowledge- and performance-testing stations, with a line of best fit plotted to the data. The unadjusted correlation between all knowledge-testing and performance-testing stations completed by junior residents was 0.66, and the unadjusted R^2 value was 0.44. The generalizability of the junior OSCE knowledge-testing stations was 0.72 and of the performance-testing stations was 0.75. The correlation, adjusted for the unreliability of the two examinations was 0.89 and the adjusted R^2 value was 0.80.

3.5 Acceptability and Educational Impact

The results of the exit survey for the October 2010 OSCE are presented in Tables 21 and 22. The residents felt they were somewhat able to demonstrate an accurate portrayal of themselves (mean rating (SD) = 5.24 (0.89)), and they felt that somewhat appropriate topics were covered (mean rating (SD) = 5.50 (1.49)), and they definitely did not feel that the time available was appropriate (mean rating (SD) = 2.41).

Residents provided the following comments on the exit surveys:

- Fairly unstructured and sounds like very different questions of all residents. Good ease though.
- Thanks for your time and effort. Well run. I could use a lot more of this.
- Timing main issues.
- Very useful, would be nice to prepare/study ahead of time (as this would be a good motivator) Would love to do this bi-annually.
- Very good, do it every 6/12 please have the Q on with the X-ray.
- Very useful exercise. Hope we do more of these before the Royal College.

The results of the exit survey for the March 2011 OSCE are presented in Tables 23 and 24. The residents felt that they were somewhat able to present an accurate portrayal of themselves (mean rating (SD) = 5.88 (0.83)). They felt that somewhat appropriate topics were covered (mean rating (SD) = 6.34 (1.19)), and felt that there was not really enough time available at each station (mean rating = 4.72 (1.58)). Only one resident provided the following comments on their exit survey:

- Much improved process over the first time
- Good content, all very good examiners

The results of the exit survey for the June 2011 OSCE are presented in Tables 25 and 26. The residents felt that they were somewhat able to present an accurate portrayal of themselves (mean rating (SD) = 5.50 (0.79)), that somewhat appropriate topics were covered (mean rating (SD) = 6.76 (0.44)), and that the time available to complete each station was somewhat appropriate (mean rating (SD) = 5.06 (1.25)).

The following comments were provided by residents on the exit surveys:

- 3 cases + 3-4 questions --> 13 pages I felt was a lot to do in 8-9 minutes
- Could have done some prep
- Peds was above my level but rest I could handle.
- Great exam. Little detail on the station on the door. Examiners were fair. Need help in peds.
- Great, thanks for organizing this!

The results of the exit survey for the September 2011 OSCE are presented in Tables 27 and 28. The residents felt that they were somewhat able to present an accurate portrayal of themselves (mean rating (SD) = 5.83 (0.88)), that somewhat appropriate topics were covered (mean rating (SD) = 6.04 (1.37)), and that the time available to complete each station was somewhat appropriate (mean rating (SD) = 5.70 (1.07)). Residents did not write any comments on the exit surveys.

Chapter 4: Discussion

This study evaluated the generalizability of four OSCEs developed by the orthopaedic surgery residency program, which took place between October 2010 and September 2011. These OSCEs were of a hybrid format, meaning that the traditional performance-testing format of the OSCE was combined with knowledge-testing components. Approximately half of the stations in each OSCE were performance-testing stations and approximately half were knowledge-testing stations. The OSCEs were constructed in this way in order to reflect the structure of the RCPSC Comprehensive Objective Examination OSCE component, for which these OSCEs were intended to prepare residents. The OSCEs were also constructed in this way to address a potential threat to feasibility, which was the limited number of faculty evaluators (needed for performance-testing stations) available on the resident half-days on which the OSCEs were administered. This study also involved construct validation measures, which determined whether residents' scores ranked as expected (i.e., whether residents of higher PGYs scored higher than residents of lower PGYs on the same OSCE) and whether the results of the knowledge-testing stations that were included correlated well with the results of performance-testing stations. This study also evaluated the examination's acceptability, feasibility, and educational impact. In assessing the utility of the OSCE in this setting, this study attempts to shed some light into whether the OSCE is an examination format that may be feasibly implemented into the setting of an orthopaedic surgery residency program, while also demonstrating good psychometric properties. There have been very few studies published on the use of the

OSCE in the setting of an orthopaedic surgery residency program, and they do not address these objectives (Beran et al., 2012; Griesser et al., 2012; Shaheen et al., 1991). In terms of evaluating the construct that the OSCE is sensitive to different levels of training, the results were inconsistent between OSCEs. The mean scores of the individual stations and the total scores in the first OSCE in October 2010 suggested that the trend was for PGY 5 residents to score better than PGY 4 residents (mean (SD) = 582.25 (55.18) and 455.12 (70.19), respectively (Figure 6, Table 1). The two-way ANOVA revealed a significant effect of year of residency (F(1,30) = 11.027, p = 0.005). The generalizability analysis indicated that 58.57% of the variance in scores was due to the year of residency. The results from this OSCE suggest that the OSCE can differentiate between increasing levels of training. This finding is supported in the literature. For example, Hodges and McIlroy (2003) conducted a ten-station OSCE for year 3 and year 4 residents at the University of Toronto, and found that on a five-point global rating scale with four subscales (empathy, coherence, verbal communication, and non-verbal expression), mean scores were found to be significantly higher for year 4 residents (Hodges and McIlroy, 2003).

However, the remaining OSCEs do not support this conclusion. Although the mean scores did increase with level of residency in the March 2011 OSCE (mean (SD) =712.56 (104.84) for PGY 4 and 632.19 (76.08) for PGY 5) and the June 2011 OSCE (mean (SD) = 652.24 (80.78) for PGY 1, 655.40 (76.12) for PGY 2, and 691.38 (88.04) for PGY 3)), the ANOVAs revealed no significant effect of year of residency for these OSCEs and for the September 2011 OSCE. The generalizability analyses revealed only small effects of level of training: only 5.09% of the variance in the March 2011 OSCE was explained by level of training, and none of the variance was explained by level of training in the June 2011 and September 2011 OSCEs.

The different results from the first OSCE to the subsequent OSCEs are striking. This may be explained, in part, by the fact that different cohorts of residents took these examinations. The PGY 4 residents who took the examination in September 2011 reflect the first cohort of residents who were selected after the admissions process became more stringent and competitive. The admissions process aimed to be more selective in accepting only those who the faculty members felt would be a good fit for the program and who would be hard-working and dedicated learners. Thus, the failure to detect a difference between PGY 4 residents and PGY 5 residents in September 2011, compared to the ability to detect this difference in October 2010, may be reflective of this selective cohort of PGY 4 residents in September 2011, that may, in fact, have been as competent as the PGY 5 residents at that time.

Another potential explanation for this is the way in which questions were created. The questions, for example for the senior OSCEs, were designed to be challenging to both PGY 4 and PGY 5 residents. Had the questions been designed to challenge PGY 4 residents but not PGY 5 residents, then there may have been a more striking difference between residents in each level of training. However, if residents of both PGYs were expected to be challenged by the examination, then it is not surprising that some examinations (March 2011, September 2011) showed no difference between levels of residency. The failure rate of PGY 4 residents (7/11, 64%) on the October 2010 OSCE, in which PGY 5 residents outperformed PGY 4 residents, suggests this examination may have actually been too challenging for the PGY 4 residents at their level of training.

Alternatively, the low sample size in each OSCE may make it difficult to detect a significant difference in performance between the different levels of training, if one truly exists. The effect sizes were large for the OSCEs in which means appeared to increase with level of training (d = -2.01 for the October 2010 OSCE, d = -0.88 for the March 2011 OSCE, $\varepsilon^2 = 0.053$ for the June 2011 OSCE). These large effect sizes suggest that a difference may truly exist, but that the study may not be powered to detect these differences.

A finding that was not accounted for in the hypotheses of this study was that the analysis of variance indicated that there was a significant main effect of station type for each of the OSCEs. Residents consistently scored higher on performance-testing stations than on knowledge-testing stations. One possible explanation for this is that knowledge-testing stations were more difficult than performance-testing stations. Another possible explanation for this finding may be leniency on the part of faculty evaluators (who were only present in performance-testing stations) in scoring residents. This is evident in the distribution of scores for the September OSCE (Figure 5). The knowledge-testing scores appear to be normally distributed, but the performance-testing scores are skewed left. However, faculty were much less lenient in evaluating residents on the OSCE than what has been demonstrated in the literature and experienced in the orthopaedic surgery residency program with other assessment tools used to evaluate the upper levels of Miller's pyramid, in particular the ITER. The distribution of the scores of each OSCE show that most residents are not grouped at the very top of the distribution (90-100%) for any of the OSCEs (Figures 3-5) and faculty were not reluctant to fail a number of residents as is the case with ITERs (Feldman et al., 2004; Finlay et al., 2006). The failure rates for the senior OSCEs are as follows: 7/17 (41%) in October 2010, 5/16 (31%) in March 2011, and 2/18 (11%) in September 2011. For the junior OSCE, the failure rate was 2/20 (10%). The high failure rate for the first two OSCEs suggests that the cutoff could have been made to be below 60%, since the aim was to have a cutoff that would result in only two or three residents failing the OSCE in total. The ability of residents to fail the OSCE but not the ITER may be explained by the fact that only one evaluator completes each ITER and has the sole responsibility of failing the resident on a particular rotation or on the FITER. However, in the OSCE, each evaluator is only responsible for failing residents on one station, and may be less reluctant to do so because of the belief that residents may perform better on other stations in the OSCE. An alternative explanation is in the nature of the OSCE as a formative rather than summative assessment tool. There are no consequences to failing the OSCE because it is a formative assessment tool. Failing the OSCE can even provide helpful feedback to a resident that he/she is scoring below their peers in performance and knowledge, and provide them the opportunity to remedy the deficit through increased studying and increased practice. The consequences to failing the ITER, however, may be not receiving credit for a particular rotation, or in the case of the FITER not advancing to attempt the RCPSC Comprehensive Objective Examination.

In terms of the evaluation of the reliability of the examination, all OSCEs achieved G coefficients over 0.70, thus achieving the standard required for high-stakes examination. The eight-station October 2010 OSCE had a G coefficient of 0.73, the nine-station June 2010 OSCE had a G coefficient of 0.82, and the ten-station March 2011 and September 2011 OSCEs had G coefficients of 0.71 and 0.87, respectively. These results indicate that McMaster University's orthopaedic surgery residency program's OSCEs have produced reliable results. In fact, the reliabilities for each OSCE far exceeded the expected reliability for a formative assessment tool and, since all are greater than 0.70, have also exceeded the expected reliability of a summative assessment tool. Since the only summative assessment tools in the program, the ITER and the FITER have demonstrated poor reliability in the literature, it is important to have a method of observing residents that can produce results that are more consistent and discriminatory.

For the June 2011 OSCE, the G coefficient for the four knowledge-testing stations was 0.72, and 0.76 for the five knowledge-testing stations in the September 2011 OSCE. The four knowledge-testing stations in the October 2010 and March 2011 OSCEs failed to achieve G coefficients of at least 0.70 (G = 0.58 and 0.68, respectively). A D study on each revealed that 7 and 5 knowledge-testing stations would be necessary to achieve a G coefficient of at least 0.70 in the March and October OSCEs, respectively. This indicates that in order to achieve a generalizable estimate of knowledge in an OSCE format, 4 to 7 stations would be necessary in this setting.

The G coefficients for the performance-testing stations in the four October 2010, five June 2011, and five September 2011 stations were 0.71, 0.75, and 0.82, respectively. The March 2011 OSCE's six performance-testing stations only achieved a G coefficient of 0.63. In order for the G coefficient to reach at least 0.70, the D study indicated that nine stations would be necessary. Therefore, 4 to 9 stations would be needed for a generalizable measure of performance to be made in this setting.

It appears that knowledge- and performance-testing stations achieved similar reliabilities. Therefore, substituting knowledge-testing stations for performance-testing stations appears to be accomplishable with no compromise to the reliability of the examination, which in fact showed to be greater than 0.70 for all the OSCEs. This is important because organizing enough faculty examiners poses a challenge to the feasibility of the OSCE, since faculty members may be occupied with other obligations, such as having to be in the operating rooms on residents' half-days when the OSCEs are administered. Including stations such as MCQs, SAQs, and spot diagnoses, which do not require an examiner to be present, can greatly improve the feasibility of the OSCE, without threatening its reliability.

To validate the decision to include knowledge-testing stations, the correlation between the two station types was determined in order to gauge the extent to which they would rank residents similarly. The correlations were analyzed separately for junior and senior residents, in case the level of training impacted this relationship. For the junior OSCE, the correlation adjusted for the unreliability of each station type was high: Pearson's r was calculated to be 0.89. The disattenuated R^2 value generated from the linear regression analysis indicated that 80% of the variance in performance-testing station scores could be explained with the knowledge-testing station scores. This relationship was even higher in the senior residents: a perfect correlation, adjusted for the average reliabilities of each station type, between knowledge-testing and performance-testing stations was found. The disattenuated R^2 value indicated that 100% of the variance in performance-testing station scores could be predicted by the variance in the knowledge-testing station scores. The knowledge-testing station results perfectly predicted the performance-testing station results. These findings suggest that not much more information regarding residents' rankings is gained in performance-testing over knowledge-testing, and challenges the implicit superiority of performance-testing Miller's pyramid.

There is further evidence to challenge the implied superiority in Miller's pyramid of evaluating performance and action. For instance, Norcini et al. (2002) found knowledge-testing to be highly predictive of actual clinical performance (the “does” section of Miller's pyramid). They compared the mortality outcomes related to myocardial infarction for internists and cardiologists board-certified with a multiple-choice examination (number of patients = 13 910), to self-designated internists and cardiologists who did not successfully complete the board examination (number of patients = 2 719) (Norcini et al., 2002). They found that the mortality rate was 19% lower for patients of board-certified physicians (Norcini et al., 2002). Another study evaluated the Physician Review Program, which involves an examination given to physicians whose competence has been called into question by peers, and includes a variety of testing formats such as MCQs, OSCEs, interactions with SPs, chart-stimulated recall, and problem-based clinical oral examinations (Davis et al., 1990). Davis et al. (1990) found that multiple choice tests better predicted the results of reassessment better than OSCEs: the correlation between the MCQ and the judgement of competence by evaluators was 0.60, whereas the correlation was 0.46 for the OSCE (Norman, 2005).

Although the results of knowledge tests have shown to be highly predictive of the results of performance tests and assessments of action, this does not mean that they actually measure the same construct. It is therefore important to select an assessment strategy based not on its apparent implied ranking on Miller's pyramid, but on the basis of the educational context, the purpose of testing, the resources available, and the attitudes of faculty and students (Norman et al., 1991; Van der Vleuten et al., 1991). In the context of the orthopaedic surgery residency program, it was necessary to implement the OSCE for a number of reasons. The OSCE was designed to help residents prepare for the RCPSC Comprehensive Objective Examination, which has an OSCE component structured like the hybrid OSCEs in McMaster University's orthopaedic surgery residency program (i.e., it combines knowledge- and performance-testing stations). It has been shown that experience with the format of a high-stakes examination can improve pass rates for that examination (Norman et al., 2010). Further, faculty members observed that senior residents, prior to the OSCE being introduced into the program, tended to be anxious about the upcoming RCPSC Comprehensive Objective Examination OSCE component. The OSCE was implemented to provide residents with experience with a hybrid OSCE examination to alleviate their concerns about the upcoming certification examination, and with the expectation that this familiarity would subsequently lead to improved pass rates for the certification examination. One resident's comment on an exit survey suggested that he/she did in fact find the examination helpful for the upcoming certification exam. They wrote, "Very useful exercise. Hope we do more of these before the Royal College."

Since it is known that assessment drives learning, the assessment of clinical and operative skills tested in performance-based stations can, theoretically, encourage the development of those skills (Norman, 2005). For example, residents may review their learning materials prior to the OSCE, in order to prepare for the knowledge-testing stations which directly test their knowledge, and for performance-testing stations which require residents to incorporate their knowledge into solving clinical problems or demonstrating technical skills. Residents may also practice discussions in which they explore case management to prepare for the performance-testing stations, many of which are of this format. Residents in the program did in fact practice these discussions in sessions with the chief resident prior to the OSCEs. This increased preparation for the OSCE may not only prepare residents for the examination and result in better overall scores on the OSCE, but the enhanced knowledge and practice in case management may benefit residents when they are managing actual clinical cases.

If faculty members would like to observe and measure how residents perform clinically and provide feedback on this to residents, the OSCE approach can be very useful. Prior to the OSCE, the only assessment based on observation of the residents' clinical performance was the ITER. However, the results from the ITER tended to not be discriminatory, with most residents scoring highly on the evaluation, as predicted by the literature (Feldman et al., 2004; Finlay et al., 2006). There was a need for a method of evaluating residents' operative and clinical skills that would produce consistent and discriminatory scores, which this study has demonstrated the OSCE can do, with G coefficients that consistently exceeded 0.70.

A limitation to this aspect of the study was that some senior residents (2010-2011) did not attend all the OSCEs, and the senior residents in the 2011-2012 academic year included a new cohort of PGY 4 students. Therefore, the residents did not all undergo the same OSCEs and the same OSCE stations. The comparisons being made between knowledge-testing and performance-testing stations were not consistently based on the same residents and stations.

Exit surveys were administered to determine the extent to which the residents found the examination to be acceptable. Residents answered the exit survey questions on a 7-point Likert scale, with 1 representing definitely not and 7 representing definitely. Most residents felt that the OSCE was an environment in which they were able to provide a somewhat authentic demonstration of their abilities: residents rated this between 5.24 and 5.88.They also felt that somewhat appropriate topics were covered, and rated this between 5.50 and 6.46 on the same scale. As residents became more familiar with the OSCE format, they also became more comfortable with the time available: senior residents rated whether the time available was appropriate 2.41 (definitely not), then 4.72 (not really), then 5.70 (somewhat). Junior residents felt that the time available was somewhat appropriate and rated this 5.06. The residents also felt that they were somewhat well-prepared for what the stations would contain by the station prompts, rating this between 5.79 and 6.37. On these domains covered on the questionnaire, it appears the examination was acceptable to the residents in the program.

The remaining exit survey questions do not fit the scale's anchors, and their results potentially may not be reflective of the examinees' actual perceptions of the examinations. The mean ratings residents provided for the level of anxiety compared to a traditional MCQ/SAQ exam ranged from 4.97 to 6.74. This was taken to mean that the OSCEs were seen as more stressful than a traditional MCQ/SAQ. This is not necessarily a shortcoming of the examination, since its intent is to prepare residents for their upcoming RCPSC certification examination, which is likely to be stressful because it is a high-stakes examination. The mean rating the residents provided for the extent to which the OSCE was more stressful than they anticipated ranged from 3.76 to 4.77, indicating that the examination likely met their expectations in terms of stress. The residents did not seem to find the examination to be very difficult, rating the difficulty between 4.44 and 5.15.

The comments that the residents provided on the exit surveys indicated satisfaction with the examination, and a motivation to continue with the assessment tool. Two residents who took the first OSCE requested that this examination be conducted biannually, and one requested more OSCE examinations in general. Seven residents commented that the examination was good, great, useful, or well run. One junior resident’s comment informed the educational impact of the OSCE. The resident identified that the OSCE made him/her aware that he/she needed help in paediatrics, suggesting that, at least for this resident in particular, the OSCE was able to provide immediate feedback on his/her performance, and demonstrate areas in which the resident needed to make improvements. Two residents identified that they could have prepared better for the OSCE, suggesting that they recognized that there may have been deficits in their knowledge, technical skills, or ability to work through clinical cases, such that they were not able to complete the OSCE with the level of preparation with which they entered the examination. This recognition can allow these residents to remedy these deficits by preparing for subsequent OSCE examinations. One of these residents suggested that the OSCE would be a “good motivator” for them to study.

A limitation to this aspect of the study was that the rating scale did not suit 3 of the 7 questions on the survey. Therefore, mean scores may not accurately reflect the perceptions residents had of the OSCE in those three domains. Further, some aspects of acceptability that were not covered may have contributed to a better understanding of the residents' impressions of the OSCE, such as the residents' impressions of fairness and educational value of the OSCE, for example (Norman et al., 1991). A revision to the scale that accounts for these limitations is recommended for future evaluations of the OSCE in this population. Additionally, few residents provided comments on the exit surveys, which were the only source of evidence to inform the educational impact of the examination.

This analysis of the first four OSCEs developed by McMaster University's orthopaedic surgery residency program found that the examinations were reliable, acceptable to residents, and could be made more feasible with the inclusion of stations that did not require evaluators. This had not previously been assessed in the literature on OSCEs in the setting of an orthopaedic surgery residency program. The expectation that the OSCE could differentiate between different levels of training could not be confirmed with these results, likely due to the low power of the small samples to detect this difference. This study found knowledge testing to be highly predictive of performance testing, supporting previous literature indicating that performance-testing is not superior in and of itself to knowledge testing. Contextual factors should be considered when determining the approach to assessing residents, and multiple forms of assessment are encouraged.

References

Al-Jarallah, K., Moussa, M.A.A., Shehab, D., & Abdella, N. (2005). Use of interaction cards to evaluate clinical performance. *Medial Teacher* **27:** 369-374.

Auewarakul, C., Downing, S.M., Pradisuwan, R. & Jaturamrong, U. (2005). Item analysis to improve reliability for an internal medicine undergraduate OSCE. *Advances in Health Sciences Education: Theory and Practice* **10:** 105-113.

Beran, M.C., Awan, H., Rowley, D., Samora, J.B., Griesser, M.J., & Bishop, J.Y. (2012). Assessment of Musculoskeletal Physical Examination Skills and Attitudes of Orthopaedic Residents. *The Journal of Bone and Joint Surgery* **94:** 1-8.

Bloch, R. & Norman, G.R. (2011). G String IV: Version 6.11: User Manual. http://fhsperd.mcmaster.ca/g_string/download/g_string_4_manual_611.pdf. Modified 30 June 2011. Accessed 20 November 2011.

Chipman, J.G., Webb, T.P., Shabahang, M., Heller, S.F., Van Camp, J.M., Waer, A.L., Luxenberg, M.G., Christenson, M. & Schmitz, C.C. (2011). A multi-institutional study of the Family Conference Objective Structured Clinical Exam: a reliable assessment of professional communication. *The American Journal of Surgery* **201:** 492-497.

Cleland, J.A., Knight, L.V., Rees, C.E. Tracey, S., & Bond, C. (2008). Is it me or is it them? Factors that influence the passing of underperforming students. *Medical Education* **42:** 800-809.

Cook, D.A., & Beckman, T.J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine* **119:** 166.e7-166.e16.

Cook, D.A., Beckman, T.J., Mandrekar, J.N., & Pankratz, V.S. (2010). Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability. *Advances in Health Sciences Education* **15:** 633-645.

Cusimano, M.D., Cohen, R., Tucker, W., Murnaghan, J., Kodama, R., & Reznick, R. (1994). A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). *Academic Medicine* **69:** 571.

Davis, D.A., Norman, G.R., Painvin, E.L., Ragbeer, M.S. & Rath, D. (1990). Attempting to ensure physician competence. (1990). *JAMA – Journal of the American Medical Association* **263:** 2041-2042.

De Lima, A.A., Conde, D., Costabel, J., Corso, J., & Van der Vleuten, C. (2011). A laboratory study on the reliability estimations of the mini-CEX. *Advances in Health Sciences Education Theory and Practice* [Epub ahead of print].

Dore, K.L., Kreuger, S., Ladhani, M., Rolfson, D., Kutz, D., Kulasegaram, K., Cullimore, A.J., Norman, G.R., Eva, K.W., Bates, S., & Reiter, H.I. (2010). The Reliability and Acceptability of the Mulitple Mini-Interview as a Selection Instrument for Postgraduate Admissions. *Academic Medicine* (suppl.) **85:** S60-S63.

Downing, S.M., & Haladyna, T.M. (2009). Validity and Its Threats. In S.M. Downing & R. Yudkowksy (Eds.), *Assessment in Health Professions Education* (21-56). New York: Taylor and Francis.

Duerson, M.C., Romrell, L.J. & Stevens, C.B. (2000). Impacting Faculty Teaching and Student Performance: Nine Years' Experience With the Objective Structured Clinical Examination. *Teaching and Learning in Medicine* **12:** 176-182.

Durning, S. J., Cation, L. J., Markert, R. J., & Pangaro, L. N. (2002). Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Academic Medicine* **77:** 900-904.

Epstein, R.M. (2007). Assessment in Medical Education. *New England Journal of Medicine.* **356:** 387-396.

Eva, K.W. (2003). On the Generality of Specificity. *Medical Education* **37:** 587-588.

Evanview, N., Holt, G., Kreuger, S., Farrokhyar, F., Petrisor, B., Dore, K., Bhandari, M., & Ghert, M. (2013). The Orthopaedic In-Training Examination: Perspectives of Program Directors and Residents from the United States and Canada. *Journal of Surgical Education* **70:** 528-536.

Feldman, L.S., Hagarty, S.E., Ghitulescu, G., Stanbridge, D., & Fried, G.M. (2004). Relationship Between Objective Assessment of Technical Skills and Subjective In-Training Evaluations in Surgical Residents. *Journal of the American College of Surgeons* **198:** 105-110.

Finlay, K., Norman, G.R., Stolberg, H., Weaver, B., & Keane, D.R. (2006). In-training evaluation using hand-held computerized clinical work sampling strategies in radiology residency. *Canadian Association of Radiologists Journal* **57:** 232-237.

Griesser, M.J., Beran, M.C., Flaigan, D.C., Quackenbush, M., Van Hoff, C., & Bishop, J.Y. (2012). Implementation of an Objective Structured Clinical Exam (OSCE) into Orthopaedic Surgery Residency Training. *Journal of Surgical Education* **69:** 180-189.

Harden, R.McG. (1988). What is an OSCE? *Medical Teacher* **10:** 19-22.

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

Hatala, R., Mar,R.S., Cuncic, C. & Bacchus, C.M. (2011). Modification of an OSCE format to enhance patient continuity in a high-stakes assessment of clinical performance. *BMC Medical Education* **11:** 23-27.

Hodges, B. & McIlroy, J.H. (2003). Analytic global OSCE ratings are sensitive to level of training. *Medical Education* **37:**1012-1016.

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

Howley, L.D. (2004). Performance Assessment in Medical Education: Where We've Been and Where We're Going. *Evaluation & the Health Professions* **27:** 285-303.

Kogan, J.R., Bellini, J.M., & Shea, J.A. (2003). Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mini-CEX) in a medicine core clerkship. *Academic Medicine* (suppl.) **10:** s33-s35.

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

Miller, G.E. (1990). The Assessment of Clinical Skills/Competence/Performance. *Academic Medicine* (suppl.) **65:** S63-S67.

Neufeld, V.R. (1985). Written Examinations. In V.R. Neufeld & G.R. Norman (eds), *Assessing Clnical Competence.* New York: Springer Publishing Company.

Norcini, J.J., Diserens, D., Day, S.C., Cebul, R.C., Schwartz, J.S., Beck, L.H., Webster, G.D., Schnabel, T.G., & Elstein, A.S. (1990). The scoring and reproduceability of an essay test of clinical judgement. *Academic Medicine* (suppl.) **65:** S41-S42.

Norcini, J.J., Lipner, R.S., & Kimball, H.R. (2002). Certifying examination performance and patient outcomes following acute myocardial infarction. *Medical Education* **36:** 853-859.

Norcini, J.J., Swanson, D.B., Grosso, L.J., & Webster, G.D. (1985). Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education* **19:** 238-247.

Norman, G.R. (2005). Editorial – Inverting the pyramid. *Advances in Health Sciences Education: Theory and Practice* **10:** 85-88.

Norman, G.R. (2002). Research in medical education: three decades of progress. *British Medical Journal* **324:** 1560-1562.

Norman, G., Neville, A., Blake, J.M., & Mueller, B. (2010). Assessment steers learning down the right road: impact of progress testing on licensing examination performance. *Medical Teacher* **32:** 496-499.

Norman, G.R. & Streiner, D.L. (eds). (2000). *Biostatistics: The Bare Essentials*. Hamilton: B.C. Decker Inc.

Norman, G.R., Tugwell, P., Feightner, J.W., Muzzin, L.J. & Jacoby, L.L.. (1985). Knowledge and clinical problem-solving. *Medical Education* **19:** 344-356.

Norman, G.R., Van der Vleuten, C.P.M. & De Graaf, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education* **25:** 119-136.

Pelgrim, E.A.M., Kramer, A.W.M., Mokkink, H.G.A., van den Elsen, L., Grol, R.P.T.M., & Van der Vleuten, C.P.M. (2011). In-training assessment using direct observation of single-patient encounters: a literature review. *Advances in Health Sciences Education* **16:** 131-142.

Petrusa, E.R., Blackwell, T.A. & Ainsworth, M.A. (1990). Reliability and Validity of an Objective Structured Clinical Examination for Assessing the Clinical Performance of Residents. *Archives of Internal Medicine* **150:** 573-577.

Reddy, S. & Vijayakumar, S. (2000). Evaluating Clinical Skills of Radiation Oncology Residents: Parts I and II. *International Journal of Cancer* **90:** 1-12.

Reznick, R., Smee, S., Rothman, A., Chalmers, A., Swanson, D., Dufresne, L., Lacombe, G., Baumer, J., Poldre, P., Lavasseur, L., Cohen, R., Mendez, J., Patey, P., Boudreau, D., Bérard, M. (1992). An Objective Structured Clinical Examination for the Licentiate: Report of the Pilot Project of the Medical Council of Canada. *Academic Medicine* **67:** 487-489.

Richards, M.L., Paukert, J.L., Downing, S.M., & Bordage, G. (2007). Reliability and Usefulness of Clinical Encounter Cards for a Third-Year Surgical Clerkship. *Journal of Surgical Research* **140:** 139-148.

Royal College of Physicians and Surgeons of Canada. CanMEDS 2005 Framework. http://rcpsc.medical.org/residency/certification/examinfo/orthopedic_e.pdf. Modified September 2010. Accessed 26 November 2011.

Royal College of Physicians and Surgeons of Canada. Format of the Comprehensive Objective Examination in Orthopaedic Surgery. http://www.royalcollege.ca/rc/faces/oracle/webcenter/portalapp/pages/viewDocument.jspx?document_id=TZTEST3RCPSCED002102&_afrLoop=3750368237468674&_afrWindowMode=0&_afrWindowId=3r7lkv4jv_1#%40%3F_afrWindowId%3D3r7lkv4jv_1%26document_id%3DTZTEST3RCPSCED002102%26_afrLoop%3D3750368237468674%26_afrWindowMode%3D0%26_adf.ctrl-state%3D3r7lkv4jv_17. Modified 2013. Accessed 3 October 2013.

Royal College of Physicians and Surgeons of Canada. Specialty Training Requirements in Orthopaedic Surgery. http://www.royalcollege.ca/cs/groups/public/documents/document/y2vk/mdaw/~edisp/tztest3rcpsced000681.pdf. Modified 2010. Accessed 3 October 2013.

Shaheen, M.A.E.K., Badr, A.A., & Al-Khudairy, N. (1991). *Sultan Quaboos University Medical Kau Univeristy Journal: Medical Sciences* **1:** 57-64.

Streiner, D.L. & Norman, G.R. (eds). (2008). *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford: Oxford University Press.

Turnbull, J. & Barneveld, C. (2002). Assessment of clinical performance: In-training evaluation. In G.R. Norman, C.P.M. Van der Vleuten, & D.I (eds), *International* *handbook of research in medical education*. Kluwer Academic Publishers, Dordrecht.

Van Dalen, J., Kerkhofs, E., Verwijnen, G.M., Van Knippenberg-van der Berg, B.W., Van den Hout, H.A., Scherpbier, A.J.J.A. & Van der Vleuten, C.P.M. (2002). Predicting communication skills with a paper-and-pencil test. *Medical Education* **36:** 148-153.

Van der Vleuten, C.P.M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education* **1:** 41-67.

Van der Vleuten, C.P.M. & Newble, D.T. (1995). How can we test clinical reasoning? *The Lancet* **345:** 1032-1034.

Van der Vleuten, C.P.M., Norman, G.R., & DeGraff, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education* **25:** 110-118.

Van der Vleuten, C.P.M. & Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine* **2:** 58-76.

Van Luijk, S.J. & Van der Vleuten, C.P.M. (1991). A comparison of checklists and rating scales in performance-based testing. In I.R. Hart & R.M. Harden (eds), *More Developments in Assessing Clinical Competence*. Can-Heal, Montreal.

Walsh, M., Bailey, P.H., & Koren, I. (2009), Objective structured clinical evaluation of clinical competence: an integrative review. *Journal of Advanced Nursing* **65:** 1584-1585.

Wass, V., Van der Vleuten, C., Shatzner, J. & Jones, R. (2001). Assessment of clinical competence. *The Lancet* **357:** 945-949.

Yudkowsky, R., Alseidi, A. & Cintron, J. Beyond fulfilling the core competencies: an objective structured clinical examination to assess communication and interpersonal skills in a surgical residency. *Current Problems in Surgery* **61:** 499-503.

Appendix 1: Minimum Training Requirements for Orthopaedic Surgery Residents, as Determined by the Royal College of Physicians and Surgeons of Canada

Five years (60 months) of approved residency training in Orthopaedic Surgery. One block of training is defined as a four (4) week rotation. This period must include:

1. Twenty six (26) blocks of foundational surgery training as a junior resident. This must follow the relevant Royal College standards.
    - 1.1. Minimum of (6) six blocks but no more than 13 blocks as a junior resident in Orthopaedic Surgery
    - 1.2. This foundational surgery training must include a minimum of (1) one block in each of the following:
        - 1.2.1. Critical care
        - 1.2.2. A service that provides initial trauma management (such as Emergency Medicine, General Surgery, trauma team, Orthopaedic Surgery, or Plastic Surgery)
        - 1.2.3. General Surgery and/or Vascular Surgery
        - 1.2.4. Internal Medicine and its relevant subspecialties
2. Thirty nine (39) blocks of further residency training in Orthopaedic Surgery
3. The entire residency program must have sufficient exposure to attain the Objectives of Training. This must include:
    - 3.1. The equivalent of at least six (6) blocks in Paediatric Orthopaedic Surgery
    - 3.2. The equivalent of at least three (3) blocks in each of the following rotations:
        - 3.2.1. Trauma

3.2.2. Sports Medicine

3.2.3. Spine Surgery

3.2.4. Oncologic Orthopaedic Surgery

3.2.5. The equivalent of at least three (3) blocks of adult reconstruction/arthroplasty in each of the following:

3.2.5.1. Upper limb

3.2.5.2. Foot and ankle

3.2.5.3. Hip and knee

3.3. The equivalent of at least two (2) blocks of training in Community Orthopaedic Surgery

3.4. At least six (6) blocks of this period must be spent as a senior resident in Orthopaedic Surgery

Appendix 2: Tables and Figures

Figure 1. Miller's Pyramid

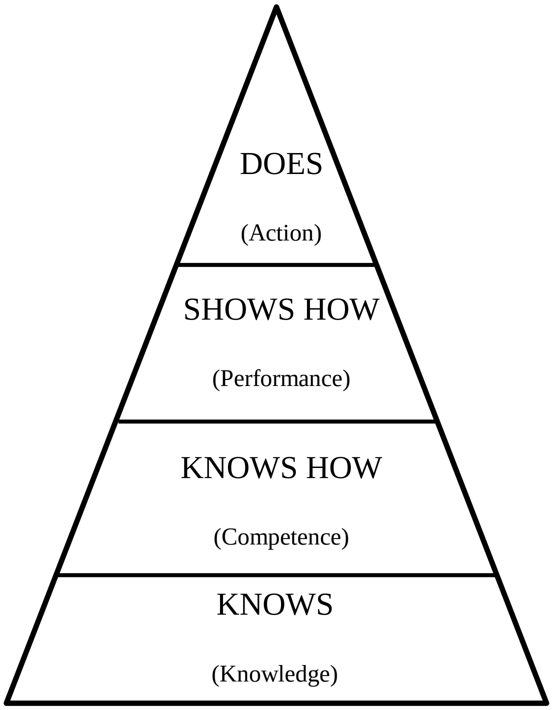

Table 1. October 2010 OSCE. The first 4 columns represent knowledge-testing stations; the next four columns represent performance-testing stations. Min. = minimum, max. = maximum, artho = arthroplasty, ST = spine trauma, peds = paediatrics.

Station	**MCQ**	**Arthro**	**Fracture**	**ST**	**Peds**	**Trauma**	**Tumour**	**Sawbone**	**Total (/800)**	**Total (/100)**
PGY 4										
Mean	39.09	63.07	60.61	38.64	66.06	61.21	61.52	64.94	**455.12**	**56.89**
SD	12.41	17.56	10.76	16.73	16.18	11.86	18.76	9.99	**70.19**	**8.77**
Min.	25.00	31.25	45.83	18.75	40.00	46.67	33.33	51.43	**336.43**	**42.05**
Max.	60.00	81.25	79.17	68.75	80.00	80.00	80.00	82.86	**542.44**	**67.81**
PGY 5										
Mean	59.17	77.08	62.50	52.60	82.22	78.89	88.58	81.19	**582.24**	**72.78**
SD	13.20	11.64	14.19	15.61	6.89	2.72	18.27	12.14	**55.18**	**6.90**
Min.	45.00	56.25	41.67	37.50	73.33	73.33	54.83	65.71	**493.46**	**61.68**
Max.	80.00	87.50	83.33	81.25	86.67	80.00	100.00	95.71	**655.60**	**81.95**
PGY 4 and 5										
Mean	46.18	68.01	61.27	43.57	71.76	67.45	71.07	70.67	**499.99**	**62.50**
SD	15.76	16.81	11.67	17.28	15.55	12.88	22.41	13.13	**89.17**	**11.15**
Min.	25.00	31.25	41.67	18.75	40.00	46.67	33.33	51.43	**336.43**	**42.05**
Max.	80.00	87.50	83.33	81.25	86.67	80.00	100.00	95.71	**655.60**	**81.95**

Table 2. March 2011 OSCE. The first four columns represent knowledge-testing stations; the next six columns represent performance-testing stations. F&A = foot and ankle, upper ex. = upper extremity.

Station	**F&A SA**	**MCQ**	**SAQ**	**SAQ**	**Technical**	**Tumour**	**Peds**	**Upper Ex.**	**Sports**	**Arthro**	**Total (/1000)**	**Total (/100)**
PGY 4												
Mean	57.25	44.32	60.25	51.25	84.13	55.63	71.25	68.75	61.88	77.50	**632.19**	**63.22**
SD	15.49	13.25	21.07	15.60	13.43	21.29	11.88	16.20	24.34	13.63	**76.08**	**7.61**
Min	40.00	27.27	34.00	38.57	60.00	30.00	45.00	40.00	30.00	60.00	**510.84**	**51.08**
Max	82.00	59.09	84.00	85.71	98.00	80.00	80.00	90.00	95.00	95.00	**725.12**	**72.51**
PGY 5												
Mean	68.75	63.64	70.25	46.43	86.00	73.13	80.00	78.75	70.00	75.63	**712.56**	**71.26**
SD	15.60	18.02	16.85	15.61	19.59	17.51	8.86	13.82	32.62	8.21	**104.84**	**10.48**
Min	48.00	40.91	42.00	24.29	40.00	40.00	70.00	60.00	10.00	60.00	**559.48**	**55.95**
Max	96.00	95.45	88.00	74.29	100.00	90.00	90.00	95.00	100.00	85.00	**881.74**	**88.17**
PGY 4 and 5												
Mean	63.00	53.98	65.25	48.84	85.06	64.38	75.63	73.75	65.94	76.56	**672.38**	**67.24**
SD	16.15	18.25	19.14	15.28	16.25	20.89	11.09	15.44	28.12	10.91	**97.74**	**9.77**
Min	40.00	27.27	34.00	24.29	40.00	30.00	45.00	40.00	10.00	60.00	**510.84**	**51.08**
Max	96.00	95.45	88.00	85.71	100.00	90.00	90.00	95.00	100.00	95.00	**881.74**	**88.17**

Table 3. June 2011 OSCE. The first four columns represent knowledge-testing stations; the next five columns represent performance-testing stations. FA = forearm.

Station	**MCQ**	**Spot Diagnosis**	**SAQ**	**SAQ**	**Technical**	**Hip**	**FA**	**Leg Pain**	**Leg Lesion**	**Total (/900)**	**Total (/100)**
PGY 1											
Mean	52.00	58.18	54.35	87.86	88.00	84.00	76.00	82.25	69.60	**652.24**	**72.47**
SD	13.04	12.61	24.93	13.27	2.74	8.94	11.40	3.17	15.37	**80.78**	**8.98**
Min.	40.00	40.91	26.09	71.43	85.00	70.00	60.00	80.25	54.00	**554.27**	**61.59**
Max.	70.00	72.73	86.96	107.14	90.00	90.00	90.00	87.50	94.00	**754.95**	**83.88**
PGY 2											
Mean	58.75	71.59	52.45	78.57	82.50	75.00	90.00	83.55	63.00	**655.40**	**72.82**
SD	18.08	22.63	23.40	21.34	7.56	7.56	10.69	5.28	3.95	**76.12**	**8.46**
Min.	40.00	45.45	17.39	42.86	65.00	60.00	70.00	76.25	58.00	**564.97**	**62.77**
Max.	90.00	100.00	89.13	107.14	90.00	80.00	100.00	91.25	69.50	**764.06**	**84.90**
PGY 3											
Mean	51.43	75.32	71.12	86.22	88.57	85.71	82.86	85.86	64.29	**691.38**	**76.82**
SD	9.00	16.57	22.55	13.43	3.78	9.76	17.04	6.29	10.16	**88.04**	**9.78**
Min.	40.00	45.45	39.13	67.86	85.00	70.00	50.00	75.00	42.50	**521.46**	**57.94**
Max.	60.00	90.91	100.00	107.14	95.00	100.00	100.00	95.00	71.50	**777.68**	**86.41**
PGY 1, 2, and 3											
Mean	54.50	69.55	59.46	83.57	86.00	81.00	84.00	84.03	65.10	**667.20**	**74.13**
SD	13.48	18.02	22.91	15.86	5.78	9.07	13.93	4.97	9.69	**75.93**	**8.44**
Min	40.00	40.91	17.39	42.86	65.00	60.00	50.00	75.00	42.50	**521.46**	**57.94**
Max	90.00	100.00	100.00	107.14	95.00	100.00	100.00	95.00	94.00	**777.68**	**86.41**

Table 4. September 2011 OSCE. The first five columns represent knowledge-testing stations; the next five columns represent performance-testing stations. UE = upper extremity.

	UE SAQ	**MCQ**	**Spot Diagnosis**	**Hip SAQ**	**Tumour SAQ**	**Lesion**	**Spine**	**Hip Trauma**	**Peds**	**MMI**	**Total (/1000)**	**Total (/100)**
PGY 4												
Mean	45.31	81.82	84.85	72.92	94.93	87.50	84.17	78.83	90.83	83.67	**804.82**	**80.48**
SD	7.84	8.13	12.08	16.61	5.94	2.74	7.36	10.23	2.04	5.85	**21.76**	**2.18**
Min.	37.50	72.73	68.18	50.00	86.96	85.00	75.00	60.00	90.00	78.00	**780.27**	**78.03**
Max.	56.25	90.91	100.00	100.00	102.17	90.00	95.00	90.00	95.00	91.00	**835.13**	**83.51**
PGY 5												
Mean	53.91	77.27	71.30	79.17	88.22	84.58	78.61	80.25	79.58	78.92	**772.03**	**77.20**
SD	21.78	17.55	19.70	17.13	15.19	16.98	23.92	12.84	23.78	10.05	**129.26**	**12.93**
Min.	18.75	36.36	27.27	50.00	65.22	55.00	15.00	48.00	30.00	62.00	**543.38**	**54.34**
Max.	87.50	100.00	100.00	100.00	104.35	100.00	95.00	95.00	95.00	93.00	**921.01**	**92.10**
PGY 4 and 5												
Mean	51.04	78.79	75.82	77.08	90.46	85.56	80.46	79.78	83.33	80.50	**782.96**	**78.30**
SD	18.50	14.95	18.37	16.74	13.05	13.81	19.83	11.74	19.93	8.99	**105.85**	**10.58**
Min.	18.75	36.36	27.27	50.00	65.22	55.00	15.00	48.00	30.00	62.00	**543.38**	**54.34**
Max.	87.50	100.00	100.00	100.00	104.35	100.00	95.00	95.00	95.00	93.00	**921.01**	**92.10**

Figure 2. Distribution of Scores for the October 2010 OSCE

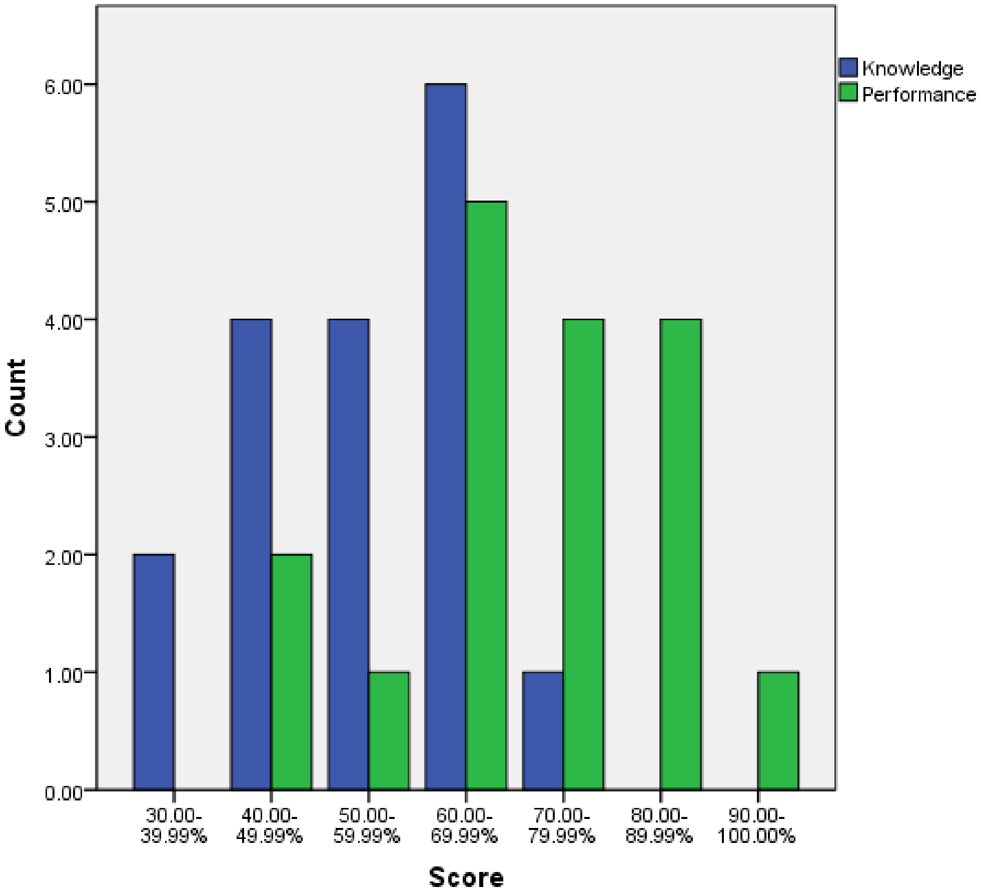

Figure 3. Distribution of Scores for the March 2011 OSCE

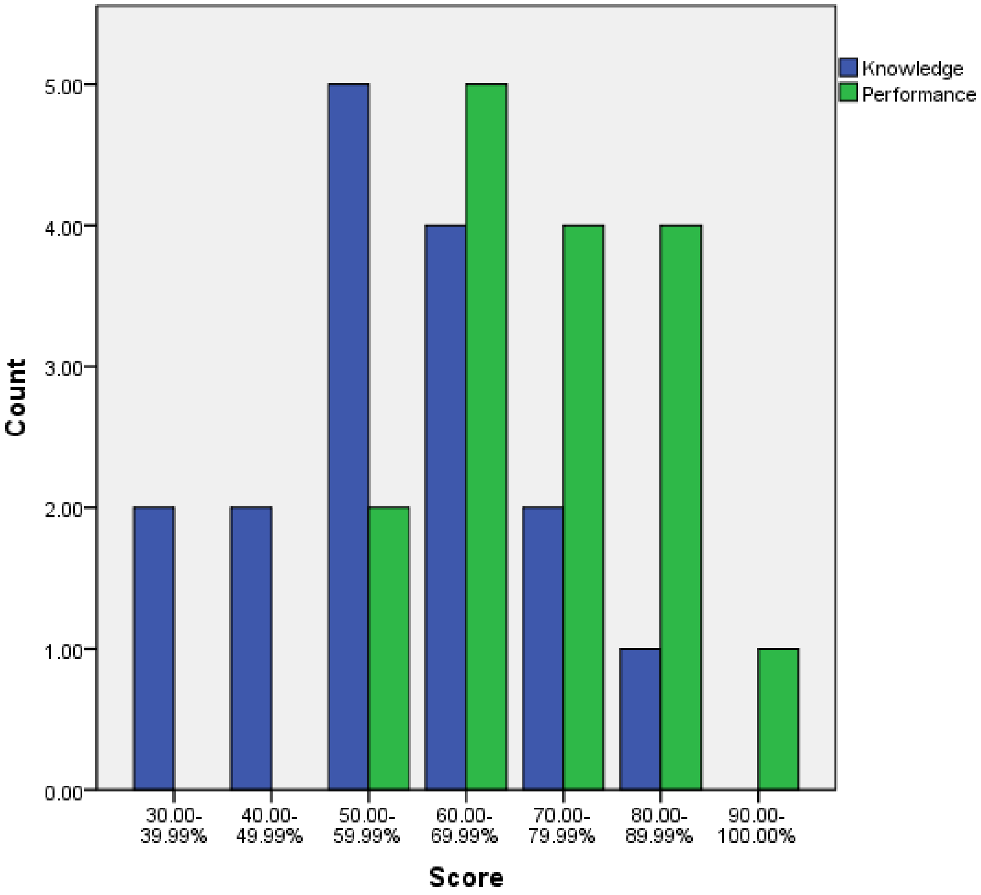

Figure 4. Distribution of Scores for the June 2011 OSCE

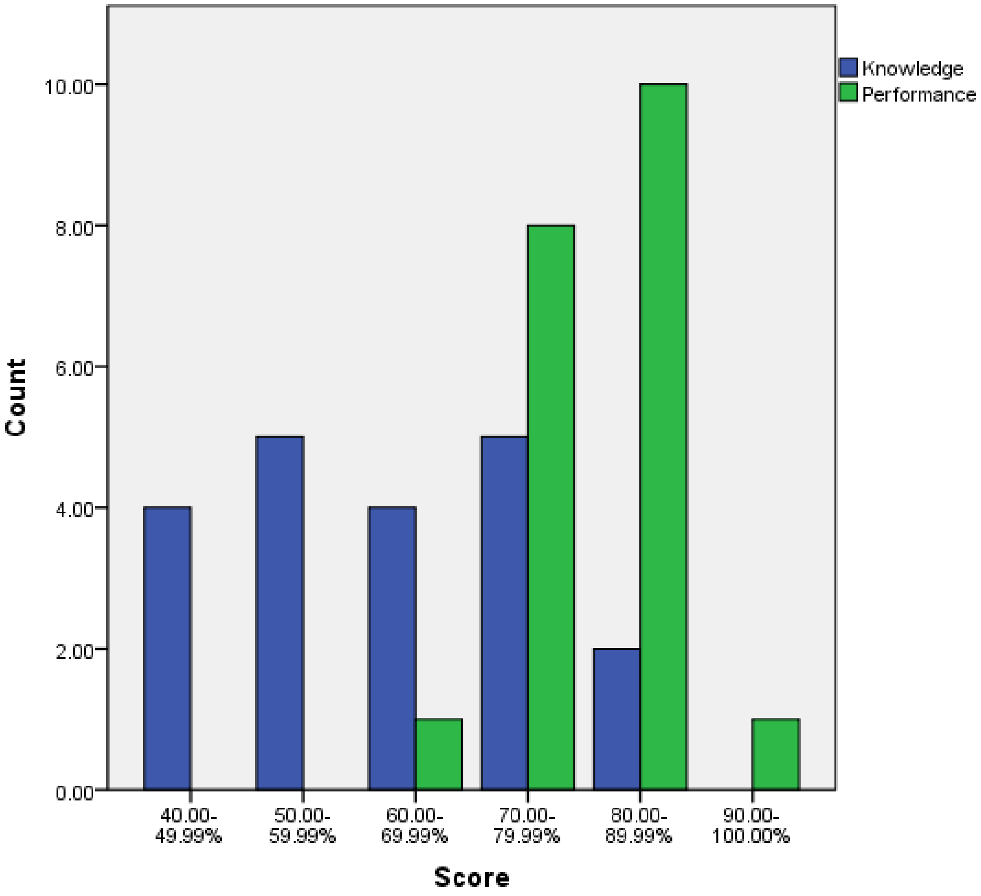

Figure 5. Distribution of Scores for the September 2011 OSCE

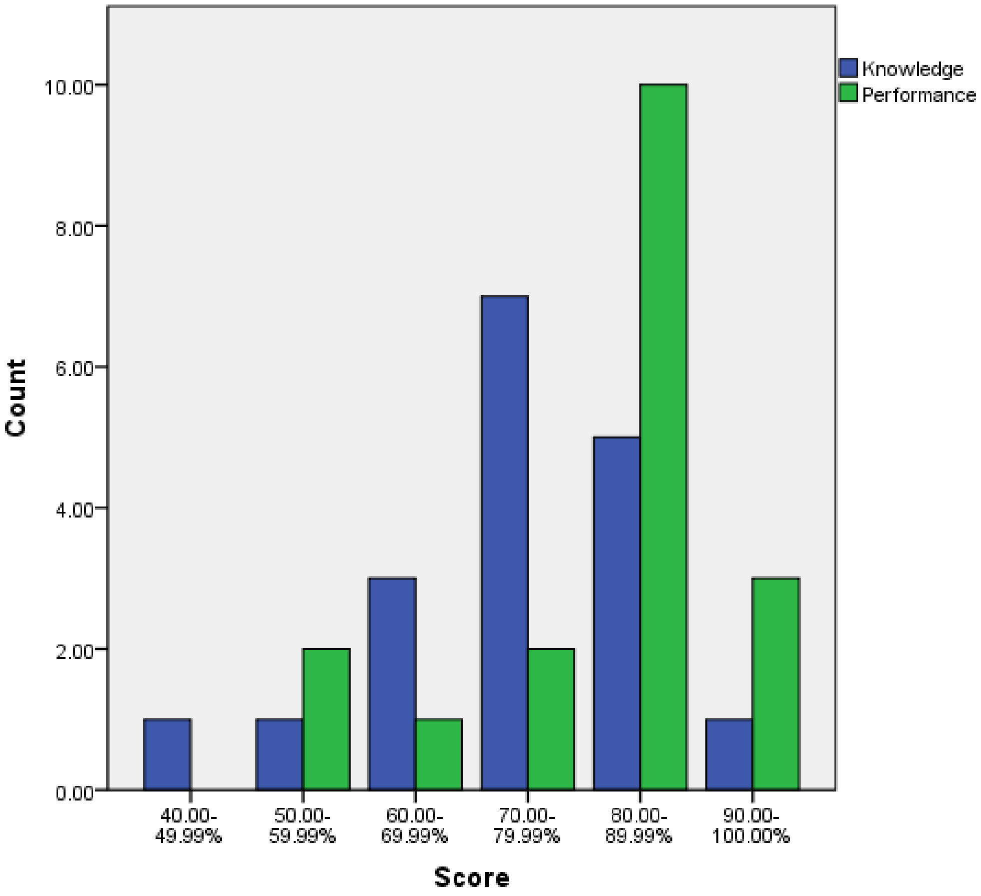

Figure 6. A Comparison of PGY 4 and PGY 5 Residents on their Mean Total OSCE Scores, Converted to a Score out of 100, in Each Senior OSCE

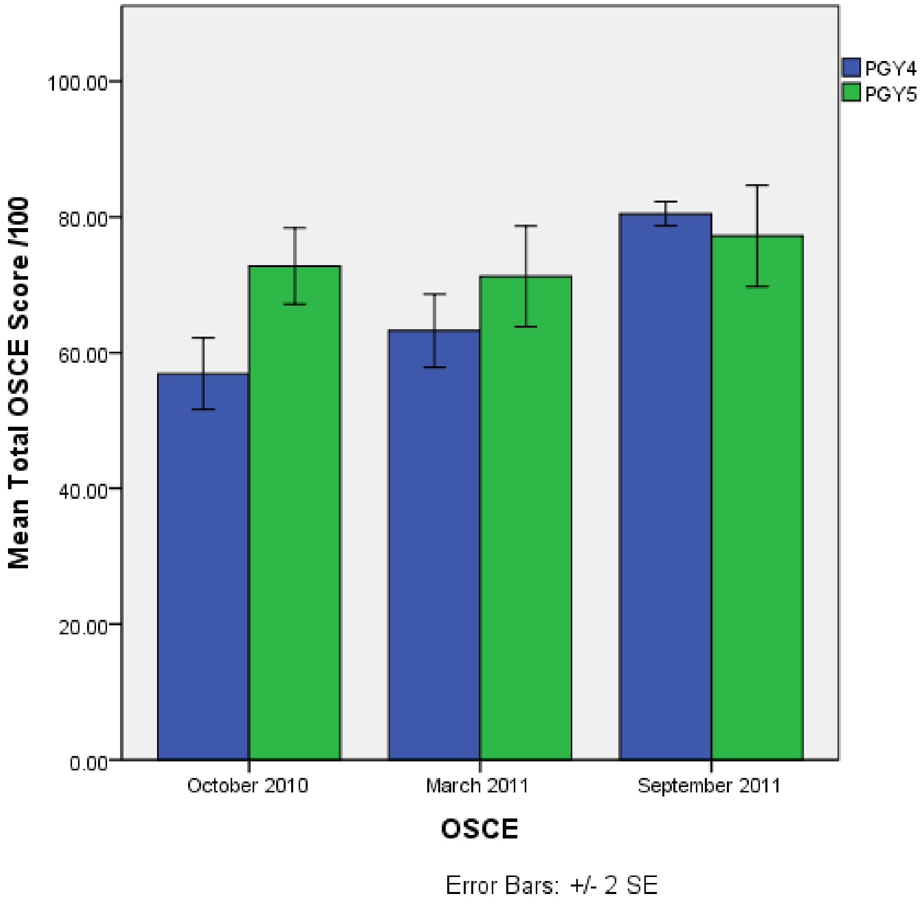

Figure 7. A Comparison of PGY 1, 2, and 3 Residents on their Mean Total OSCE Scores, converted to a Score out of 100, in the Junior OSCE

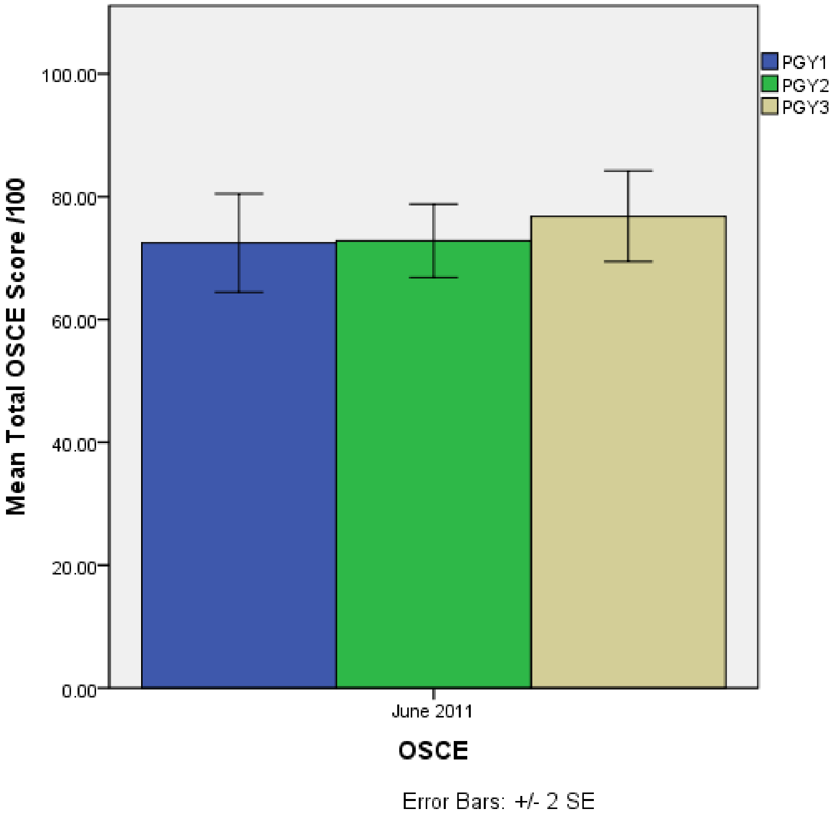

Figure 8. Mean Total Scores Separated by Station Type in the October 2010 OSCE

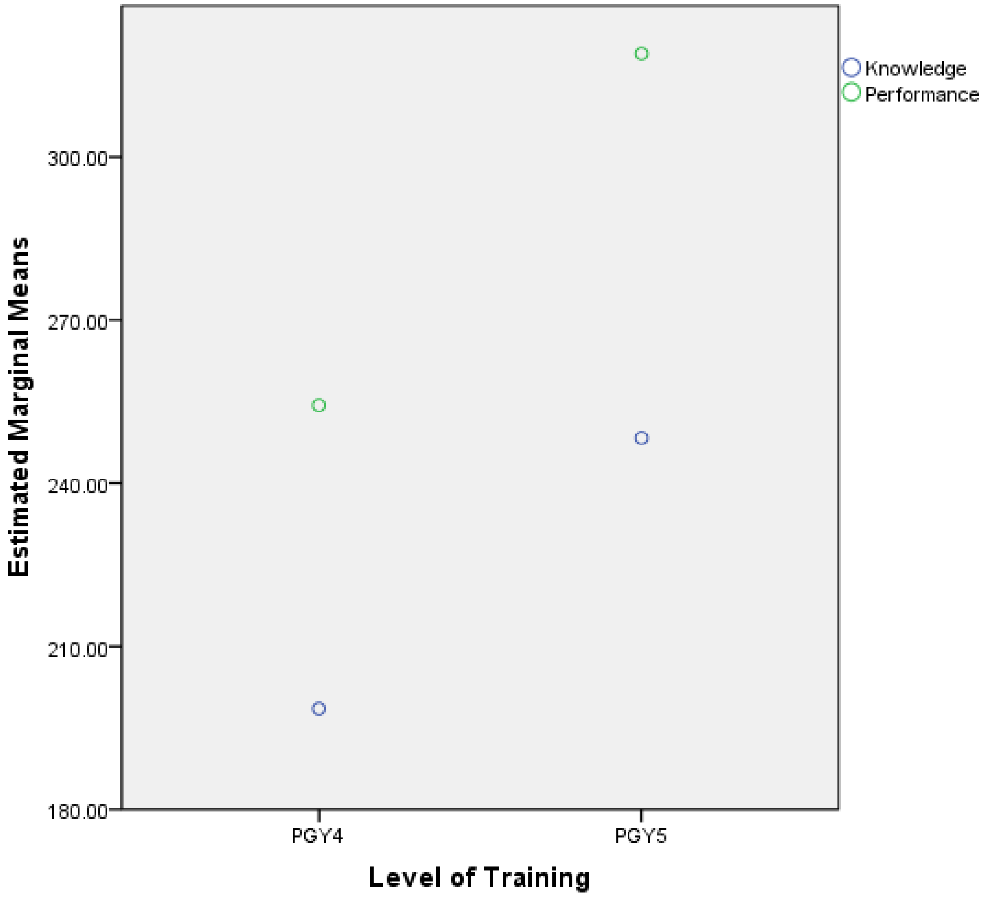

Figure 9. Mean Total Scores Separated by Station Type in the March 2011 OSCE

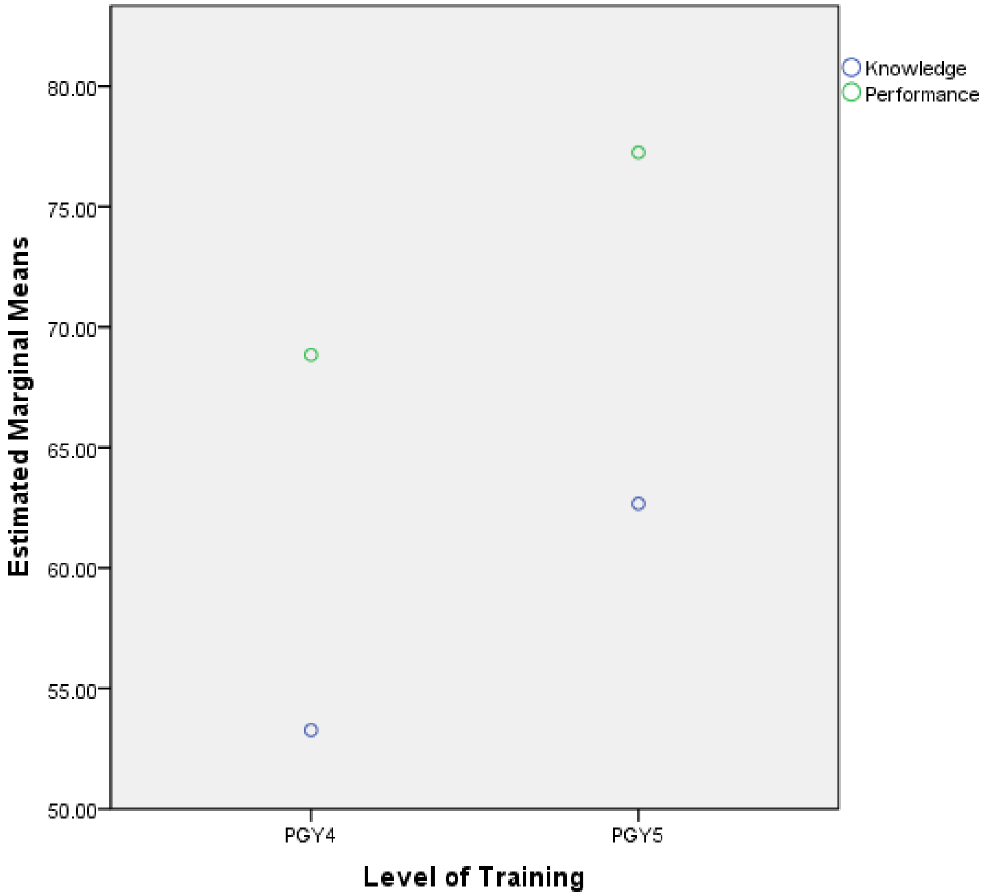

Figure 10. Mean Total Scores Separated by Station Type in the June 2011 OSCE

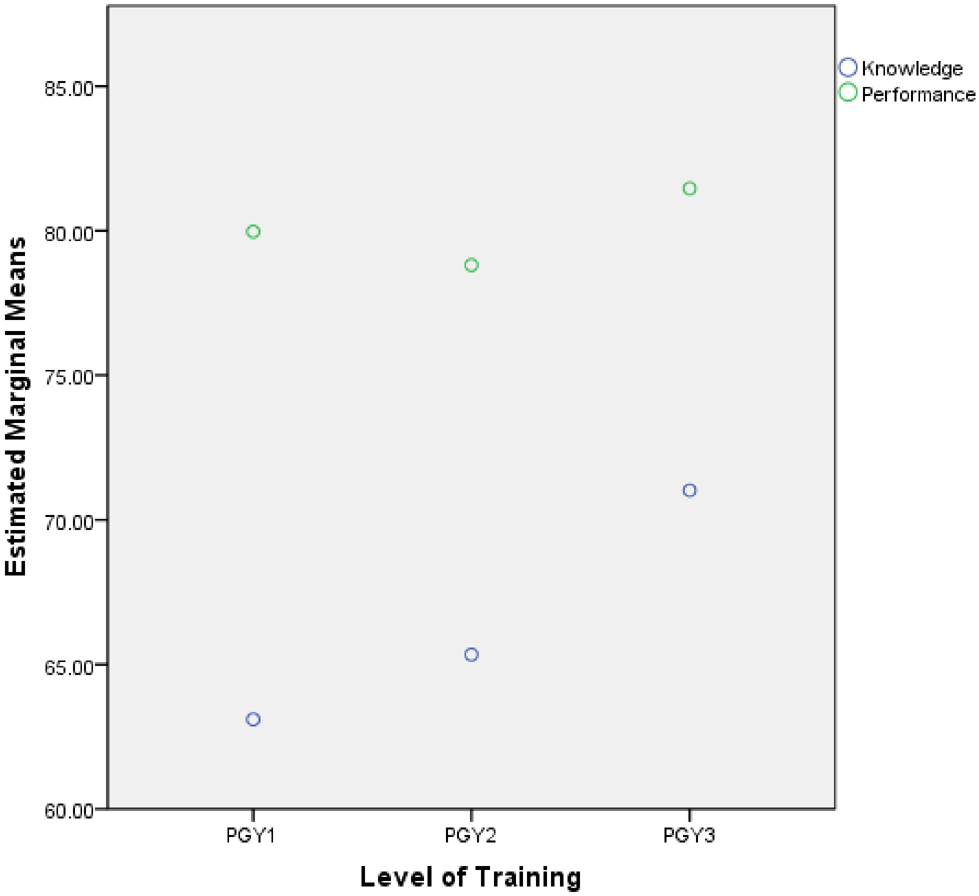

Figure 11. Mean Total Scores Separated by Station Type in the September 2011 OSCE

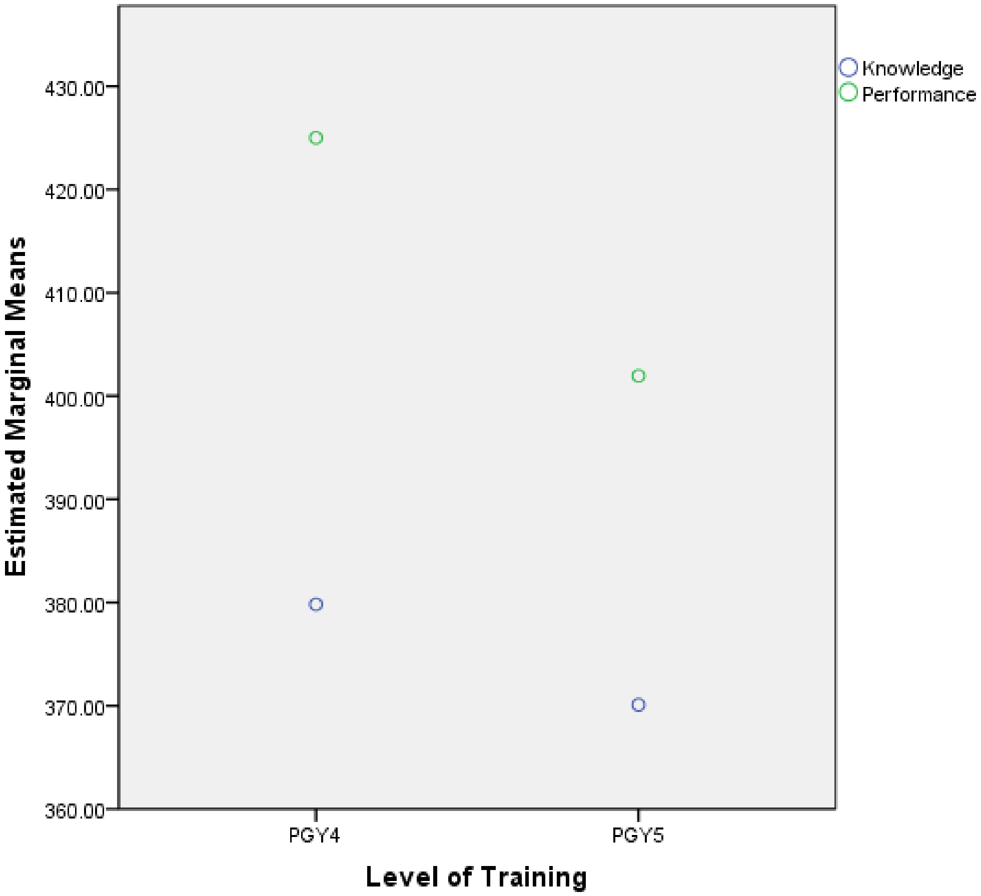

Figure 12. Correlation of Knowledge-Testing and Performance-Testing Stations for all Senior Residents

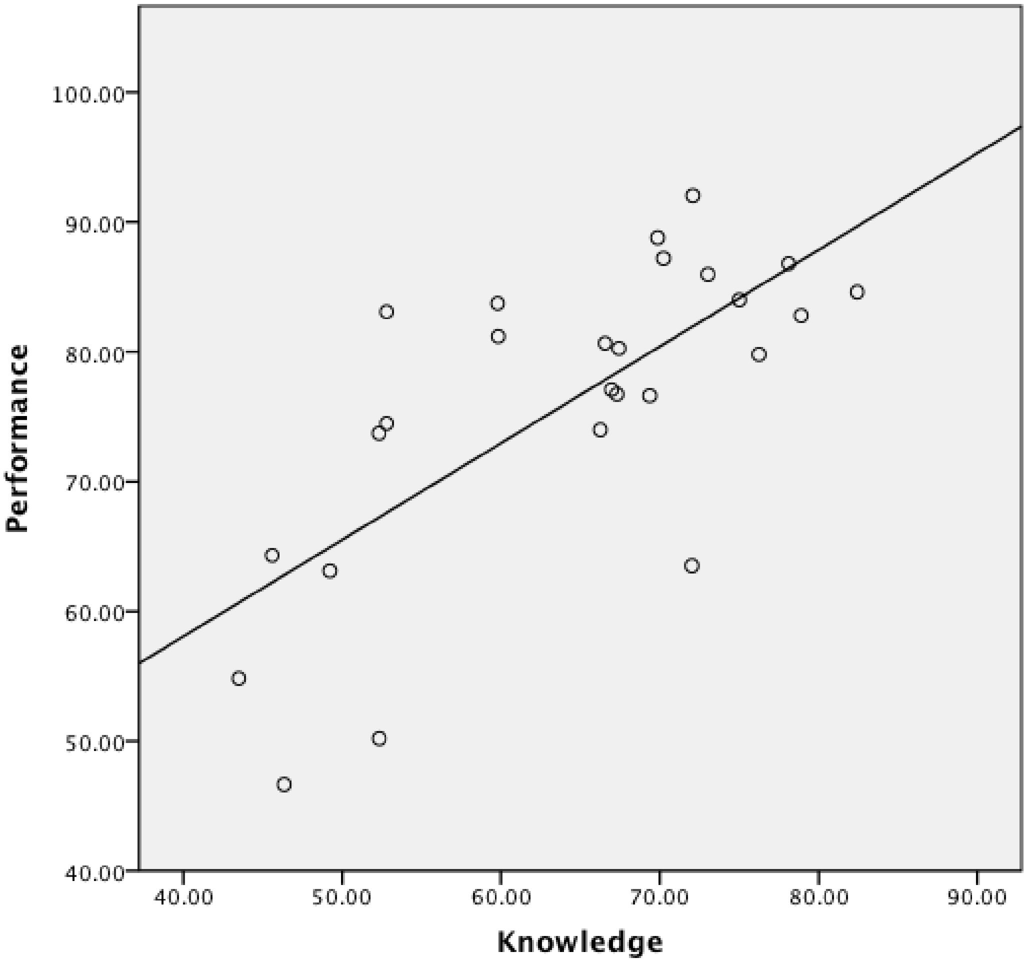

Figure 13. Correlation of Knowledge-Testing and Performance-Testing Stations for all Junior Residents

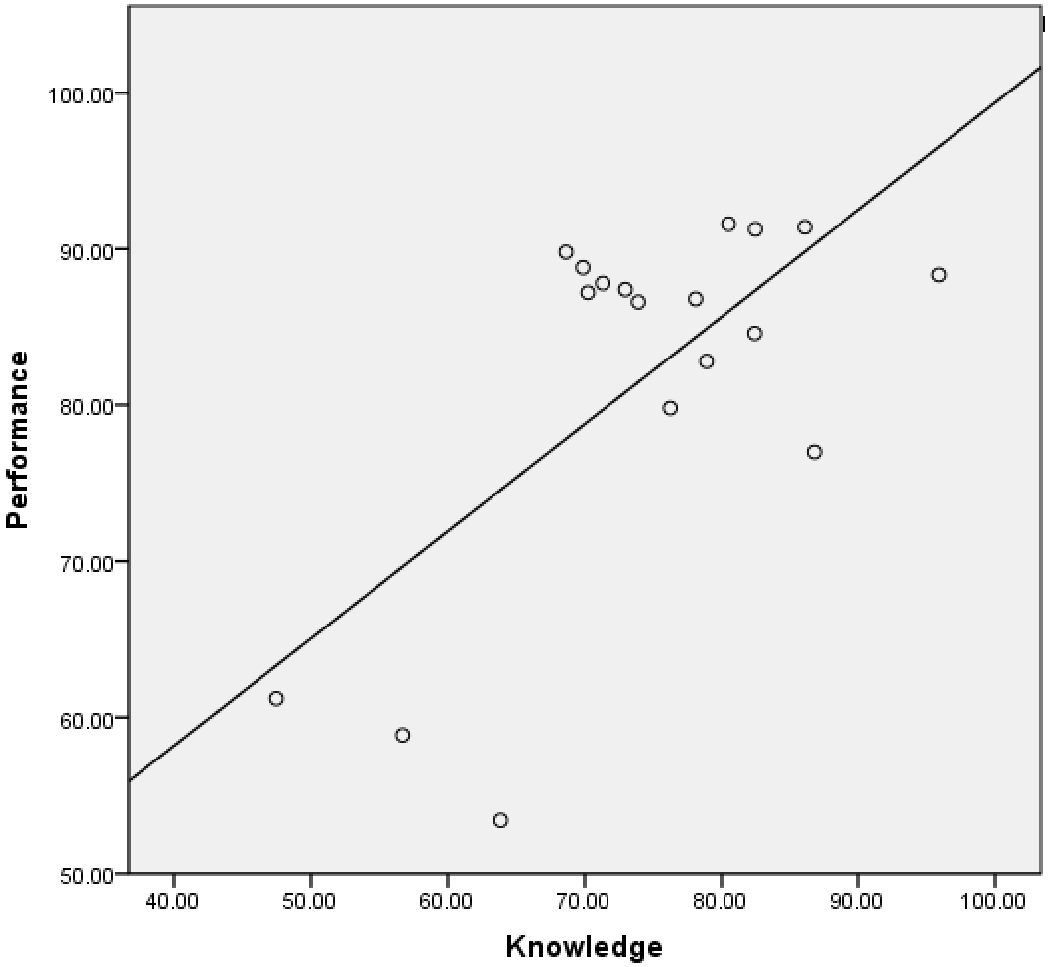

Table 5. Variance Components for the October 2010 OSCE

Effect	Variance Component	Proportion of Variance
Y	116.843	58.57%
P:Y	48.720	24.42%
S	14.716	7.38%
YS	0.738	0.37%
PS:Y	18.467	9.26%

Table 6. Variance Components for the October 2010 OSCE Knowledge-Testing Stations

Effect	Variance Component	Proportion of Variance
Y	63.98582	15.83%
P:Y	52.24729	12.93%
S	124.96030	30.92%
YS	9.26855	2.29%
PS:Y	153.64770	38.02%

Table 7. Variance Components for the October 2010 OSCE Performance-Testing Stations

Effect	Variance Component	Proportion of Variance
Y	173.52531	48.13%
P:Y	70.72132	19.62%
S	-2.47967	0.00%
YS	-1.29091	0.00%
PS:Y	116.29477	32.26%

Table 8. Variance Components for the March 2011 OSCE

Effect	Variance Component	Proportion of Variance
Y	21.89486	5.09%
P:Y	59.29853	13.78%
S	103.21738	23.98%
YS	-0.83226	0.00%
PS:Y	245.97955	57.15%

Table 9. Variance Components for the March 2011 OSCE Knowledge-Testing Stations

Effect	Variance Component	Proportion of Variance
Y	16.32685	4.63%
P:Y	91.79592	26.02%
S	33.76297	9.57%
YS	27.89383	7.91%
PS:Y	183.02649	51.88%

Table 10. Variance Components for the March 2011 OSCE Performance-Testing Stations

Effect	Variance Component	Proportion of Variance
Y	14.67974	3.80%
P:Y	71.00045	18.37%
S	46.23457	11.96%
YS	-9.05592*	0.00%
PS:Y	254.58284	65.87%

Table 11. Variance Components for the June 2011 OSCE

Effect	Variance Component	Proportion of Variance
Y	-8.27263*	0.00%
P:Y	67.39147	19.21%
S	135.08970	38.51%
YS	15.17577	4.33%
PS:Y	133.13262	37.95%

Table 12. Variance Components for the June 2011 OSCE Knowledge-Testing Stations

Effect	Variance Component	Proportion of Variance
Y	-18.54410*	0.00%
P:Y	137.98528	26.71%
S	146.53834	28.36%
YS	22.11368	4.28%
PS:Y	210.01757	40.65%

Table 13. Variance Components for the June 2011 OSCE Performance-Testing Stations

Effect	Variance Component	Proportion of Variance
Y	-7.79946*	0.00%
P:Y	31.15021	18.98%
S	64.25288	39.14%
YS	17.36944	10.58%
PS:Y	51.39087	31.30%

Table 14. Variance Components for the September 2011 OSCE

Effect	Variance Component	Proportion of Variance
Y	-9.64148*	0.00%
P:Y	100.61191	27.89%
S	98.36760	27.26%
YS	4.42459	1.23%
PS:Y	157.38882	43.62%

Table 15. Variance Components for the September 2011 OSCE Knowledge-Testing Stations

Effect	Variance Component	Proportion of Variance
Y	-19.47078*	0.00%
P:Y	102.53237	21.51%
S	185.56823	38.93%
YS	23.41238	4.91%
PS:Y	165.13528	34.65%

Table 16. Variance Components for the September 2011 OSCE Performance-Testing Stations

Effect	**Variance Component**	**Proportion of Variance**
Y	-6.12236*	0.00%
P:Y	117.19190	47.38%
S	1.92058	0.78%
YS	-5.50803*	0.00%
PS:Y	128.25614	51.85%

Table 17. G Coefficients for all OSCEs and their Knowledge-Testing and Performance-Testing Stations

OSCE	**All Stations**		**Knowledge-Testing Stations**		**Performance-Testing Stations**	
	Number of Stations	**G Coefficient**	**Number of Stations**	**G Coefficient**	**Number of Stations**	**G Coefficient**
October 2010	8	0.73	4	0.58	4	0.71
March 2011	10	0.71	4	0.67	6	0.63
June 2011	9	0.82	4	0.72	5	0.75
September 2011	10	0.87	5	0.76	5	0.82

Table 18. D Study for the October 2010 OSCE Knowledge-Testing Stations

Number of Stations	**G Coefficient**
4	0.58
5	0.63
6	0.67
7	0.70

Table 19. D Study for the March 2011 Knowledge-Testing Stations

Number of Stations	**G Coefficient**
4	0.67
5	0.72

Table 20. D Study for the March 2011 Performance-Testing Stations

Number of Stations	**G Coefficient**
6	0.63
7	0.66
8	0.69
9	0.72

Table 21. Exit surveys for the October 2010 OSCE, Part I

	Question			
	1. Ability to Present Accurate Portrayal	**4. Extent to Which Stems on Doors Prepared for Station**	**5. Extent to Which Appropriate Topics were Covered**	**7. Appropriateness of Time Available**
Mean	5.24	5.79	5.50	2.41
SD	0.89	1.22	1.49	1.59

Table 22. Exit surveys for the October 2010 OSCE, Part II

	Question		
	2. More Anxiety than Traditional MCQ/SAQ	**3. Extent to Which Process was more Stressful than Expected**	**6. Extent to Which Stations were Difficult**
Mean	6.74	3.88	4.44
SD	0.44	1.59	0.75

Table 23. Exit surveys for the March 2011 OSCE, Part I

	Question			
	1. Ability to Present Accurate Portrayal	**4. Extent to Which Stems on Doors Prepared for Station**	**5. Extent to Which Appropriate Topics were Covered**	**7. Appropriateness of Time Available**
Mean	5.88	6.06	6.34	4.72
SD	0.83	1.33	1.19	1.58

Table 24. Exit surveys for the March 2011 OSCE, Part II

	Question		
	2. More Anxiety than Traditional MCQ/SAQ	**3. Extent to Which Process was more Stressful than Expected**	**6. Extent to Which Stations were Difficult**
Mean	5.19	4.41	5.09
SD	1.57	1.27	1.04

Table 25. Exit surveys for the June 2011 OSCE, Part I

	Question			
	1. Ability to Present Accurate Portrayal	**4. Extent to Which Stems on Doors Prepared for Station**	**5. Extent to Which Appropriate Topics were Covered**	**7. Appropriateness of Time Available**
Mean	5.50	6.26	6.76	5.06
SD	0.79	1.15	0.44	1.25

Table 26. Exit surveys for the June 2011 OSCE, Part II

	Question		
	2. More Anxiety than Traditional MCQ/SAQ	**3. Extent to Which Process was more Stressful than Expected**	**6. Extent to Which Stations were Difficult**
Mean	4.97	3.76	5.15
SD	1.32	1.43	0.70

Table 27. Exit surveys for the September 2011 OSCE, Part I

	Question			
	1. Ability to Present Accurate Portrayal	**4. Extent to Which Stems on Doors Prepared for Station**	**5. Extent to Which Appropriate Topics were Covered**	**7. Appropriateness of Time Available**
Mean	5.83	6.37	6.04	5.70
SD	0.88	0.90	1.37	1.07

Table 28. Exit surveys for the September 2011 OSCE, Part II

	Question		
	2. More Anxiety than Traditional MCQ/SAQ	**3. Extent to Which Process was more Stressful than Expected**	**6. Extent to Which Stations were Difficult**
Mean	5.17	4.77	4.82
SD	1.33	1.43	0.91

Appendix 3: Examples of OSCE Questions

Example of a multiple choice question:

Which of the following structures represents a site of compression of the median nerve at the elbow?

1- Ligament of Struthers

2- Intermuscular septum

3- Osborne's ligament

4- Fascia of the flexor carpi ulnaris

5- Flexor-pronator aponeurosis in the forearm

Example of a short answer question:

The following xrays are of a 48 year old female. She has been having increasing pain over her 1st MTP and a prominent bunion.

[Radiographs presented on next slide]

What are the radiographic abnormalities on this xray?

How would you treat this patient who wishes surgical intervention?

Example of an Oral Station Scenario:

- 41 year old male involved in an MVA. He was the driver. Seen and assessed by the trauma service. Diagnosed with a splenic laceration, liver laceration and injury to his right ankle and foot. General surgery is observing the injuries to the abdomen.

- Obvious open injury and exposed bone and joint surface. Distal neurovascular status is intact. PMHx nil, smokes 1 pack/day

Example of an MMI question:

You are on a new rotation. The staff, who is responsible for grading your performance, is discussing a previous resident in very offensive and ridiculing terms that focus on that person's ethnicity. Your 'significant other' is of that ethnic group. To make matters worse, an obligatory social function that includes partners is planned for the coming weekend. What do you do?

After the resident answers the question, the following talking points are available for discussion:

- what are the options

- what are the possible ramifications of confronting staff

-what are possible outcomes of ignoring outburst

- prompt the applicant to come up with a plan of action.

Appendix 4: Scoring of Performance-Testing Stations

October 2010 OSCE

The following adjectival scale was used in scoring residents on their performance:

1	2	3	4	5
Fails to meet expectations	Meets expectations – marginal	Meets expectations – satisfactory progress	Meets expectations - fully	Exceeds expectations

The oral stations had three categories on which this scale was completed

1. Communication;
2. Strength of knowledge; and
3. Overall performance.

The same scale was used to score the technical station, except in the following seven domains:

1. Instrument handling;
2. Fixation construct;
3. Efficiency of operation;
4. Quality of reduction;
5. Approach used and which neuromuscular plane;

6. Quality of reduction; and
7. Overall surgical competence.

March 2011and June 2011 OSCEs

For all oral stations and the technical station, residents were given a global rating out of 100 for their overall performance.

September 2011 OSCE

Residents were scored with a global rating out of 25 for oral stations in each of four categories:

1. History and physical;
2. Diagnosis;
3. Treatment plan; and
4. Description of approach.

For the MMI station, two categories were scored out of 25:

1. Communication; and
2. Content.