Designing an Assessment Program for the Competency-Based Era

by

Dr Tim Dwyer

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Institute of Medical Science
University of Toronto

© Copyright by Tim Dwyer 2016
Designing an Assessment Program for the Competency-based Era

Tim Dwyer

Doctor of Philosophy

Institute of Medical Science
University of Toronto

2016
Abstract

Designing an Assessment Program for the Competency-based Era

Doctor of Philosophy, 2016

Dr Tim Dwyer MBBS FRACS FRCSC

Institute of Medical Science, University of Toronto

Background

The introduction of CBME at the University of Toronto division of orthopedics is an opportunity to examine the concepts of CBME. While much has been written about CBME’s potential advantages and disadvantages, further research is required to support this new model of training.

Aims

The first aim was to develop and evaluate a program of performance-based, in-training assessment tools, covering the breadth of competencies including clinical skills, the intrinsic CanMEDS roles, and performance of technical procedures. The second aim was to use these assessments to examine the premise that all residents can achieve a minimal level of competency after a subspecialty rotation.
Results

The sports medicine OSCE and OSATS were used to assess clinical skills and performance of technical procedures respectively. Each of these assessment tools demonstrated a significant difference in performance between junior and senior residents, with many junior residents unable to achieve a minimal level of competence. As a result, a modified Angoff method of standard setting was used to set different pass marks for junior and senior residents when assessing clinical skills. Using simulation to assess resident performance of Entrustable Professional Activities, many junior residents were not able to achieve a minimal level of competency, despite having been previously deemed competent. Finally, an OSCE used to assess residents’ application of the intrinsic CanMEDS roles demonstrated a significant difference in performance between resident groups depending on year of training.

Conclusion

The assessment tools used to assess clinical skills, intrinsic CanMEDS roles, and performance of technical procedures demonstrated sufficient evidence of validity for use in CBME. These assessments demonstrate that many junior residents are not able to achieve a minimal level of competence. These findings affect the fundamental concept of achieving competence before progression in CBME, and lend credibility to hybrid versions of CBME.
Acknowledgments

Many, many thanks to

Brian Hodges, my supervisor, who gave me confidence,

Charlotte Ringsted, who helped me to understand what a thesis was, and what it should be,

Mahan Kulasegaram, for his insights and encouragement,

and,

Veronica Wadey, for her kindness

The research studies of this thesis were supported in part by grants from the OMeGA Medical Grants Association
This thesis is dedicated to

To my darling Megan, and my wonderful children Sebastian and Simone, who give me everything, and

To my mother Julie and father Terry, who gave me everything
# Table of Contents

**ABSTRACT** ................................................................. iii

**CHAPTER 1: INTRODUCTION** ........................................... 1

**CHAPTER 2: LITERATURE REVIEW** ................................... 3

2.1 Definitions .................................................................. 3
  2.1.1 Assessment & Evaluation .................................... 3
  2.1.2 Competence & Competencies ............................... 4
  2.1.3 Models of Expertise ......................................... 6
  2.1.4 Conclusion .................................................... 9

2.2 Competency-based Medical Education ......................... 10
  2.2.1 Background .................................................. 10
  2.2.2 Milestones ................................................... 12
  2.2.3 Entrustable Professional Activities ..................... 14
  2.2.4 Proposed advantages of CBME .......................... 16
  2.2.5 Potential disadvantages and problems .............. 19
  2.2.6 Summary of the evidence regarding CBME to date 22
  2.2.7 The Toronto experience .................................. 23
  2.2.8 Conclusion .................................................. 26

2.3 Psychometric and Assessment Theory .......................... 27
  2.3.1 Validity ....................................................... 27
  2.3.2 Reliability ................................................... 31
  2.3.3 Feasibility ................................................... 37
  2.3.4 Acceptability ............................................... 38
  2.3.5 Educational impact ....................................... 39
  2.3.6 Conclusion .................................................. 39

2.4 The Assessment of Competence ................................... 40
  2.4.1 Introduction ................................................. 40
  2.4.2 Medical Expert - Clinical Skills ....................... 44
  2.4.3 Intrinsic CanMEDS roles ................................ 44
  2.4.4 Medical Expert - Technical procedures ............ 46
  2.4.5 Objective Structured Clinical Examinations (OSCEs) 52
  2.4.6 Objective Structured Assessment of Technical Skills (OSATS) 65

2.5 Summary ............................................................... 68

**CHAPTER 3: AIMS AND HYPOTHESES** ......................... 71

**CHAPTER 4: Competency-Based Medical Education: Can Both Junior and Senior Residents Achieve Competence After a Sports Medicine Training Module?** 73

*Published by the Journal of Bone and Joint Surgery America, 2015 Dec 2 97(23):1985-91*

T Dwyer, S Wright, KM Kulasegaram, J Theodoropoulos, J Chahal, D Wasserstein, C Ringsted, B Hodges, D Ogilvie-Harris
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>How to set the bar in Competency-Based Medical Education: Standard setting using an Objective Structured Clinical Examination (OSCE)</td>
<td>87</td>
</tr>
<tr>
<td>6</td>
<td>How to assess Communication, Professionalism, Collaboration and the other Intrinsic CanMEDS Roles in Orthopaedic Residents: Use of an Objective Structured Clinical Assessment (OSCE)</td>
<td>102</td>
</tr>
<tr>
<td>7</td>
<td>Competency in Technical Procedures: Use of an Objective Structured Assessment of Technical Skill (OSATS) after a Sports Medicine Rotation</td>
<td>120</td>
</tr>
<tr>
<td>8</td>
<td>Cognitive and Psychomotor Entrustable Professional Activities: Can Simulators Help Assess Competency in Trainees?</td>
<td>136</td>
</tr>
<tr>
<td>9</td>
<td>GENERAL DISCUSSION</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>9.1 Summary of findings</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>9.2 Implications of these findings as they relate to CBME</td>
<td>182</td>
</tr>
<tr>
<td></td>
<td>9.3 Limitations</td>
<td>195</td>
</tr>
<tr>
<td>10</td>
<td>CONCLUSIONS</td>
<td>206</td>
</tr>
<tr>
<td>11</td>
<td>FUTURE DIRECTIONS</td>
<td>207</td>
</tr>
<tr>
<td>A</td>
<td>Appendix A. Curriculum map for Competency-based Curriculum at the University of Toronto Division of Orthopedics</td>
<td>214</td>
</tr>
<tr>
<td>B</td>
<td>Appendix B: Curriculum overview map, listing the intrinsic CanMEDS roles mapped to each rotation</td>
<td>215</td>
</tr>
<tr>
<td>C</td>
<td>Appendix C: Assessing Competence of Orthopaedic Residents: The Reliability and Validity of an Objective Structured Clinical Examination After a Sports Medicine Rotation</td>
<td>216</td>
</tr>
<tr>
<td>D</td>
<td>Appendix D: Simulation of Anterior Cruciate Ligament Reconstruction (ACLR) in a Dry Model</td>
<td>235</td>
</tr>
</tbody>
</table>
List of Tables

CHAPTER 4
Table 1. Station metrics and Cronbach’s Alpha if Station Removed for checklist scores and overall global rating for each of the stations. SD – standard deviation, SLAP – superior labrum anterior posterior.

CHAPTER 5
Table 2. Modified Angoff Method. Displayed are the expected checklist percentage correct by examiner type and by resident group (junior – PGY1-3, senior PGY 4&5). For all three groups of judges, a significant difference was seen both overall and for each station between junior and senior residents (p<0.001). No significant difference was seen between each group of judges. PGY – postgraduate year.

Table 3. Pass marks using modified Angoff for the junior residents, Borderline Groups (BG) and Borderline Regression (BLR) method. A significant difference was seen for the modified Angoff and the BG / BLR methods for junior residents for all stations and overall (all p<0.001).

Table 4. Pass marks using modified Angoff for the senior residents, Borderline Groups (BG) and Borderline Regression (BLR) method. No significant difference was seen on any station between the modified Angoff method and the BG / BLR methods for senior residents. n.s. – non significant.

Table 5. Number of failures established by each standard setting method, for each station and overall.

CHAPTER 6
Table 6. Breakdown of each individual station by primary and secondary Roles, and the key competencies tested, taken from the CanMEDS 2005 Physician Competency Framework. Communicator was also assessed in stations 1 and 3.

Table 7. Resident demographics.

Table 8. Total test scores and overall rating of performance.

Table 9. Internal consistency for the four Role scores with more than one rating scale.

Table 10. Correlation between the Program Directors (PD) Ratings of resident ability in each of the Intrinsic Roles, and the corresponding OSCE Role Score.
CHAPTER 7
Table 11. List of technical procedures that residents were expected to demonstrate a minimal level of competence at after a sports rotation. ACL – anterior cruciate ligament.

Table 12. Number and percentage of participants deemed competent or better in each knee station using the overall global rating scale (novice, advanced beginner, competent, proficient, expert).

Table 13. Number and percentage of participants deemed competent or better in each shoulder station using the overall global rating scale (novice, advanced beginner, competent, proficient, expert).

CHAPTER 8
Table 14. List of the top ten Entrustable Professional Activities at our institution, as determined by faculty.

Table 15. The three Entrustable Professional Activities, with a breakdown of the individual components assessed.

Table 16. Cronbach’s Alpha if Item deleted for each station. For no station was this number greater than the overall reliability, indicating that each station was performing well.

Table 17. Mean global rating scale (Standard Deviation) for each of the EPA station components. N.s. – non significant.

Table 18. The number of residents deemed competent or better at each station of each entrustable professional activity. PGY = postgraduate year; IM = intramedullary

CHAPTER 9
Table 19. Evidence of validity collected for the clinical skills OSCE.

Table 20. Evidence of validity collected for the CanMEDS OSCE.

Table 21. Results of the CanMEDS study, demonstrating the number of residents in each group able to achieve a minimal level of competence on 4/6 stations.

Table 22. Evidence of validity collected for the OSATS study.

Table 23. Results of the OSATS study, demonstrating the number of participants in each group able to achieve a minimal level of competence on 4/6 stations.
Table 24. Evidence of validity collected for the EPA study.

Table 25. The number of residents in each group able to achieve a minimal level of competence in all three stations of each EPA.

**APPENDIX C**

Table 26. Number of Residents in each Post-Graduate Year (PGY)

Table 27. Group (n=43) mean scores and standard deviation for each of six OSCE stations

Table 28. Individual station parameters

Table 29. Inter-item reliability of sub-scores and effect of year of training on domains of performance.

Table 30. Correlations (Pearson’s r) between OSCE sub-scores and OITE scores for 35 participants. Eight PGY1 residents had not written an OITE at the time of this study, so data was unavailable for correlation. ** Correlation is significant at the 0.01 level (2-tailed) * Correlation is significant at the 0.05 level (2-tailed).

**APPENDIX D**

Table 31. Task-specific checklist for the drilling of the femoral tunnel, using an anteromedial technique, for an 8 mm hamstring graft with fixation using a 20 mm loop Endobutton.

Table 32. Task-specific checklist for the drilling of the tibial tunnel.

Table 33. Task-specific checklist for passage of hamstring graft and fixation, using Endobutton fixation on the femoral side, and an interference screw on the tibial side.

Table 34. Number of participants in the study and their year of training. Three different methods of determining a minimal level of competency are shown. PGY – postgraduate year of training. GRS – global rating scale.

Table 35. Details of the sports medicine experience and exposure of each group of participants in this study.
List of Figures

CHAPTER 4

Figure 1. Box plots for (A) overall global rating, and (B) total checklist score. A significant difference was seen between both junior and senior residents, and junior residents and fellows for both (p<0.01). No difference was seen between senior residents and fellows for either.

Figure 2. Box plots for each of the domains of clinical skills. A significant difference was seen between junior and senior residents for every skill. No difference was seen between senior residents and fellows.

CHAPTER 5

Figure 3. Comparison of cut-scores using the modified Angoff, Borderline Group (BG), and Borderline Regression (BLR) methods for junior and senior residents for each station.

CHAPTER 6

Figure 4. Total station scores (% correct) by PGY for each of the stations. Each station showed a significant difference by PGY (p<0.05) except for station 3 (NAI / Non accidental injury) (p=0.068). Error bars represent standard error of the mean. PGY – postgraduate year.

Figure 5. Total test scores (% correct) for each of the CanMEDS Roles by PGY. ANOVA testing showed significant differences for all roles (communicator, collaborator, manager, profession p<0.001; advocate, scholar p<0.05). PGY – postgraduate year.

CHAPTER 7

Figure 6. (A) The Arthroscopic SOFT Shoulder Model (Arthrex, Naples, USA) bench top simulator used for insertion of glenoid anchor and labral repair. Antero-superior and anterior-inferior cannulas have been inserted into the rotator interval prior to performance of the technical procedure. (B) Instrumentation available for performance of this skill (Smith & Nephew, Andover, MA). (C) The Arthroscopic SOFT Shoulder Model bench top simulator and instrumentation used for insertion of rotator cuff anchor and passage of rotator cuff suture. Again, two cannulas have been inserted into the subacromial space prior to the OSATS.

Figure 7. (A) The knee model (Sawbones, Washington, United States) used for performance of partial medial meniscectomy, drilling of the tibial tunnel, and drilling of the femoral tunnel using both the anteromedial and transtibial technique. Again all portals were made prior to the assessment. (B) Set up provided for whipstitch of hamstring graft.
Figure 8. (A) Box plot for the overall checklist scores calculated as a percentage. (B) Box plot for the total ASSET global rating. (C) Box plot for the overall global rating. There was a significant difference by year of training for each (p<0.001).

Figure 9. Boxplots for the total ASSET score for each of the stations. A significant difference by year of training was seen for all technical skills.

CHAPTER 8

Figure 10. Technical procedures performed as a component of each entrustable professional activity are shown. (A) ORIF of oblique fibular fracture. (B) Insertion of dynamic hip screw under image guidance. (C) Performance of distal femoral cut and AP cuts for TKA. ORIF = open reduction and internal fixation.

APPENDIX C

Figure 11. (A) Box plot of total checklist score (y) versus postgraduate year of training (x). Scores on total checklists were significantly higher for PGY5 residents than PGY1, 2 and 3 residents (p<0.05); no significant difference was seen between PGY4 and PGY5 (Figure 2). (B) Box plot of global rating 1 to 5 (x) versus postgraduate year of training (y). A significant difference (p<0.0001) was seen between PGY1 and PGY5 residents. ° and * represent group outliers.

Figure 12. Box plots of total checklist scores compared with the postgraduate year of training for each of the six domains. There was no significant difference in the domain of history (p = 0.993) (Fig. 2-A), but a significant difference was seen in the domains of examination (p = 0.001) (Fig. 2-B), image interpretation (p < 0.001) (Fig. 2-C), clinical decision-making (p < 0.001) (Fig. 2-D), consent (p = 0.019) (Fig. 2-E), and surgical technique (p < 0.001) (Fig. 2-F). The degree symbol and the asterisk represent group outliers.

Figure 13. Box plots of total checklist scores compared with the number of sports rotations for each of the six domains. Again, there was no significant difference in history (p = 0.281) (Fig. 3-A), but a significant difference was seen in the domains of examination (p = 0.005) (Fig. 3-B), image interpretation (p = 0.001) (Fig. 3-C), clinical decision-making (p = 0.001) (Fig. 3-D), consent (p = 0.01) (Fig. 3-E), and surgical technique (p = 0.001) (Fig. 3-F). The degree symbol represents group outliers.

APPENDIX D

Figure 14. A: Equipment provided for ACL reconstruction in Sawbones model. B: View of the Sawbones model, with the accessory anteromedial portal being used in conjunction with an offset guide to drill the femoral tunnel. C: Drilling of the tibial tunnel. D: Passage of the femoral beath pin using an offset guide.
Figure 15. A: Boxplot for the total ASSET score for junior residents (postgraduate year 1-3), senior residents (postgraduate year 4&5), fellows, and staff surgeons. A significant difference was seen between junior and senior residents, senior residents and fellows, and fellows and staff surgeons. B: Boxplots for the total checklist score for the four groups of participants. A significant difference was also seen between junior and senior residents, as well as senior residents and fellows – no significant difference was seen between fellows and staff surgeons.

Figure 16. Boxplot for the total ASSET score for all participants, based upon blinded video review analysis. Again, a significant different was seen between participants based upon year of training for the total ASSET score (p<0.001), as well as between all groups.
List of Appendices

**APPENDIX A**: The core curriculum / modules of the Competency-based Curriculum

**APPENDIX B**: Curriculum overview map, listing the intrinsic CanMEDS Roles mapped to each rotation

**APPENDIX C**: Assessing Competence of Orthopaedic Residents: The Reliability and Validity of an Objective Structured Clinical Examination After a Sports Medicine Rotation
Journal of Bone and Joint Surgery America, 2013 Nov 20;95(22):e177
T Dwyer, J Theodoropoulos, J Herold, P Henry, D Wasserstein, M Murnaghan, V Wadey, B Hodges, J Semple, D Ogilvie-Harris

**APPENDIX D**: Simulation of Anterior Cruciate Ligament Reconstruction (ACLR) in a Dry Model
T Dwyer, J Slade Shantz, J Chahal, D Wasserstein, R Schachar, KM Kulasegaram, J Theodoropoulos, R Greben, D Ogilvie-Harris
Chapter 1
Introduction

The traditional format of postgraduate medical education is time-based, whereby residents are exposed to a fixed period of clinical experience (‘tea-steeping’) (Hodges 2010). The primary determination of graduation is thereby length of time in the training program, typically coupled, in the example of orthopedics, with an exit examination in countries such as Canada, the UK, Australia and New Zealand.

While this method of medical education has been used for over 100 years, there has been an increasing push for a move away from a time-based model, to one based upon demonstrable outcomes, as evidenced by growing debate at medical education conferences, and in medical education journals (Frank, Mungroo et al. 2010). The rationale for this move is discussed extensively in the literature review, but is primarily focused on ensuring that all residents are deemed ready for the workforce, after having demonstrated competency in the critical tasks and skills that make up their profession.

Asking residents to demonstrate competency prior to graduation is a sound concept at face value, but many issues arise in the implementation of competency-based medical education (CBME). The logistics of managing postgraduate training programs that are based solely upon the learner are complex; how to best assess resident competency in each of the medical specialties is unclear. At this time, there is little in the way of research that has emerged from CBME training programs.

In 2009, the University of Toronto Division of Orthopedics began a trial of a competency-based curriculum. In this new approach to postgraduate medical training, the curriculum became modular, with residents expected to demonstrate competency in one rotation before moving to the next. The pilot study was deemed a success, and in 2013 all incoming orthopedic residents were enrolled into the program, one of the first iterations of CBME in medical training.
The competency-based program at the University of Toronto offers an incredible opportunity to delve into the issues that surround this revolution in training, issues including what it means to be competent, how competency is assessed, and importantly, how competency develops. The literature review for this thesis will examine the existing literature surrounding these issues, and provide a rationale for studying the CBME program at the University of Toronto, with a focus on the development and evaluation of an in-training assessment program, and the attainment of competence.

Section 1 of this literature review will discuss critical terms in CBME, including the understanding of what it means to be competent or to be an expert. This section will also explore various models of expertise, and theories of clinical reasoning, in order to understand how the progression towards competence occurs during medical training. A clear understanding of these issues will be important in order to study CBME.

Section 2 focuses on CMBE, addressing concepts of milestones and entrustable professional activities as they relate to the attainment of competence. This section explores the background for the shift from traditional time-based training to one based upon outcomes, as well as the potential advantages and disadvantages of an outcomes-based approach to postgraduate education. Section 2 finishes by reviewing what research has told us thus far about CBME, and by discussing the University of Toronto experience.

Section 3 discusses psychometric theory, contrasting classical and contemporary concepts of validity, as they pertain to in-training assessment in CBME. Following this, Section 4 compares and contrasts the different assessment tools available for use in an outcomes-based program, focusing on those that can be used to assess clinical skills, the application of the Intrinsic CanMEDS Roles and the performance of technical procedures. Finally, section 5 summarizes the literature review, as it pertains to the aims of this thesis.
Chapter 2

Literature Review

2.1 DEFINITIONS

In order to discuss the role of CBME in postgraduate medical education, it is important that key terms such as assessment and evaluation are clarified. Furthermore, before competence can be demanded of residents, what competence means should be clearly defined. Terms such as competent, competence, competency, and competencies are often used interchangeably in the literature, but have subtle and important differences. It will also be important to understanding models of expertise when implementing and evaluating CBME, as these theories guide what can, and cannot be expected of medical residents as they progress through an outcome-based education program.

2.2.1 ASSESSMENT AND EVALUATION

Assessment is described as the process of collecting information or evidence of a learner’s progress over a period of time (Palomba 1999). In this way, assessment is a process that examines or measures student learning, typically as a result of an academic program, and determines whether or not intended outcomes are being achieved (Gagne 1998). This information is then used to improve both teaching and learning (Palomba 1999, Borden 2001).

Evaluation is a term used to describe the determination of the level of quality, or the process of making an overall judgment. Evaluation uses assessment information to support decisions on maintaining or changing instructional practices (Gagne 1998). Evaluation produces a global view of achievement, and focuses only on the level of quality, without any focus on why that level was attained (Palomba 1999, Borden 2001).
2.1.2 COMPETENCE / COMPETENCIES

Without doubt, the definition of competence varies – the following are examples taken from the literature, from respected and prominent educators. Norman describes competency as relating to an individual’s ability to perform in the workplace to the required standard (Norman 2000). McConnell defined competence as the knowledge, and the capacity to perform skills (McConnell 2001). Frank et al. defined competent as ‘possessing the required abilities in all domains in a certain context at a defined stage of medical education or practice’, and competence as ‘the array of abilities across multiple domains or aspects of physician performance in a certain context’ (Frank, Snell et al. 2010). Hall et al. described competence as the demonstrated ability to safely care for patients in a thoughtful and knowledgeable manner whilst maintaining acceptable standards of professional behavior (Hall, Crebbin et al. 2004). Focusing on concepts of entrustment and the ability to perform independently, ten Cate et al. defined the competent resident as being able to carry out an activity without supervision (ten Cate, Snell et al. 2010).

We don’t say that someone is competent, but rather that someone is competent at something, usually a defined task or set of tasks (Brooks 2009). Simple, and elegantly, in 2009 Brooks defined being competent as a minimum standard – a practitioner has been judged capable of performing a task (Brooks 2009). Typically, such a task is a time-bound work activity, performed by someone under standardized conditions using standardized tools (Nickols 2011). Skill refers to the ability to perform a task.

One aspect of competence that is accepted by most authors is that of context. Clearly, the determination or establishment of competence, in any domain, is dependent upon expert opinion, experience and consensus, as well as the setting (Lurie 2012). The requirements for an allocation of ‘competence’ depends on context; educators need to define a set of roles that professionals should play, with competence becoming the degree to which individuals are required to fulfill them (Hodges 2012).

Frank and colleagues discuss the concept of ‘progression of competence’, a dynamic process that develops or recedes over time (Frank, Snell et al. 2010). Because
competence depends on context, a resident might be competent to perform some, but not all procedures independently (Frank, Snell et al. 2010). Furthermore, professional competence is multifactorial – it includes academic competence, operational competence, and meta-competencies (communication, self-development, problem solving) (Talbot 2004).

**Competencies** are typically framed in prescriptive and organization terms that describe the critical components of a training program (Lurie 2012). A literature review in 2010 identified at least 173 sets of published and specific competencies for various specialties (Frank, Mungroo et al. 2010). A behavioral approach to competencies has been recommended, whereby competencies are treated as though they are clusters of work related behaviors (Catano 2007).

The most common are the ACGME core competencies and the CanMEDS framework. The six **ACGME competencies** (patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, systems-based practice) are domains in which a physician must ultimately demonstrate competence (ACGME 1999, Carraccio, Wolfsthal et al. 2002). The **CanMEDS Physician Competency Framework** has seven Roles (Medical Expert, Communicator, Collaborator, Manager, Health Advocate, Scholar, and Professional) that describe the knowledge, skills and abilities that specialist physicians require for better patient outcomes (Frank 2005). Methods of describing and assessing individual competencies are described further below.

Clearly, the whole of competence is greater than the its individual parts, with the demonstration of competency in the delivery of patient care described as meta-competency (Fleming 1993, Grant 1999, Iobst, Sherbino et al. 2010). Regardless, the assessment of competence requires valid and reliable multi-dimensional assessment (Iobst, Sherbino et al. 2010).
2.1.3 MODELS OF EXPERTISE

Most educators accept that the attainment of competence is only the first step on the pathway to clinical excellence – for this reason, considering what differentiates the expert medical practitioner from the competent one becomes important in the context of CBME. In medicine, experts are often labeled as those who have completed clinical training and have spent a reasonable time in clinical practice – this definition ignores the crucial aspect of performance (Elstein A 1990, Mylopoulos 2009). Expertise is neither authority nor experience (Alderson 2010). Furthermore, competence is not a binary concept (i.e. competent and non-competent); it is likely that there is a graduated progression towards a minimal level of competence over time that is both trainee and task dependent.

The concept of expertise is widely embraced, but poorly defined in surgery (Alderson 2010). Expertise is thought to have many characteristics, including aspects of pattern recognition, knowledge, skill, flexibility, metacognitive monitoring, teaching abilities, and available cognitive space (Dunphy and Williamson 2004). Experts are able to recognize key features of a problem more rapidly than other practitioners – perceptual expertise allows them to be highly selective in their search, and to solve routine problems without exploring many alternatives (Gobet 2005). Experts are superior at monitoring and evaluating their own performance, and are easily able to both detect and correct errors (Ericsson and Lehmann 1996).

Dreyfus and Dreyfus (1986,1996) defined the expert as being distinguished by high levels of procedural knowledge and skill (knowing how), as well as contextual flexibility (knowing when and what) (Dunphy and Williamson 2004). In the field of surgery, experts have been defined as “experienced surgeons with consistently better outcomes than non experts,” with research revealing that expert surgeons demonstrate superior performance in diagnostic and operative skills compared with non experts (Schaverien 2010). Expert surgeons typically have developed domain specific memory skills, allowing the automatization of some tasks – it is thought that this ability to carry out a task with minimal attentional effort is slow to develop (Schmidt, Norman et al. 1990, Ericsson KA 1995, Schaverien 2010).
An expert also has the capacity to remain aware of the whole situation and reflect on alternative possibilities should a decision need to be made – this metacognitive monitoring of ongoing actions is a feature of expertise, and made possible by not having to direct limited cognitive resources to more mundane aspects of tasks that have become automated (i.e. procedural steps) (Garofalo 1986). Like competence, expertise is domain, context and time specific – for example, a surgeon may be an expert at open surgery, but a novice at endoscopic surgery (Alderson 2010).

It is also important to recognize that there are two types of experts – routine experts who are skilled at routine operations, and adaptive experts, who grow in expertise because of continued engagement in problem solving (Schaverien 2010). In this way, many surgeons can become routine experts within a limited, routine range, with good results when practicing within stable conditions (Alderson 2010). This concept is relevant in CBME – by clearly defining list of important technical procedures, complemented by appropriate assessment methods, residents can achieve not just a minimal level of competency, but begin on the path to becoming a routine expert in their chosen field.

The majority of the models of expertise presented describe a series of graduated, developmental steps, with a progression towards a stage of competence or expertise (Alderson 2010). An example is the two-dimensional matrix of cognizance and competence developed by Howell (Howell 1982). In this model, learners progress through four stages; unconscious incompetence, conscious incompetence, conscious competence, and unconscious competence. In the final stage, as the task is mastered, performance occurs automatically.

Dreyfus and Dreyfus described five stages of skill acquisition (novice, advanced beginner, competent, proficient, expert), based upon small group reflections on driving a car and flying an airplane (Alderson 2010). Using these distinctions inherent in different levels of skillfulness, a novice has been described as being able to evaluate factual knowledge, with the advanced beginner able to evaluate and integrate information (Chambers and Glassman 1997). The competent practitioner is defined as being able to function individually, in a realistic work setting, while the proficient
practitioner has reached skilled status, and can apply rules to untested situations (Chambers and Glassman 1997). The expert has intuitive grasp of situations, an analytic approach to novel situations, and is visionary (Dreyfus 1986). It is important to realize that a learner will commonly be at different stages depending on the skill set, and the particular task (Carraccio, Benson et al. 2008).

There is some evidence that the stages of skills acquisition are represented by a curve (ten Cate, Snell et al. 2010), as demonstrated in studies assessing the acquisition of manual skills in anesthesiology (Konrad, Schupfer et al. 1998), colonoscopy (Tassios, Ladas et al. 1999), breast biopsy (Liberman, Benton et al. 2001), and internal medicine (Hayashino, Fukuhara et al. 2006). Benner et al. studied theories of experiential (trial and error) learning in the field of nursing, finding that nurses typically took a year to progress from a novice to an advanced beginner, then typically another 1.5 to 2 years to become competent (Benner 1996). In this study, the competent practitioner was distinguished by increased organizational ability and technical skill, performing well in familiar situations but had limited ability to integrate domains such as clinical knowledge. Each of these studies reveals a gradual transition from stage to stage.

Theories have been proposed with regards the process by which learners progress through these stages, in regards to both clinical reasoning, and the acquisition of psychomotor skills. Schmidt et al. described four stages of clinical reasoning (Schmidt, Norman et al. 1990). In stage 1, learners develop casual networks i.e. the linkage of signs and symptoms to the pathophysiology of disease (Schmidt, Norman et al. 1990, Carraccio, Benson et al. 2008). In stage 2, learners develop higher order thinking after exposure to patients, synthesizing the networks developed in stage 1, allowing them to begin to practice solving clinical problems. Illness scripts characterize stage 3 - after increased experience and exposure, learners begin to be able to recognize diagnostic categories – pattern recognition. In stage 4, physicians unconsciously use illness scripts, and begin to use these stored scripts to solve less common and more complex problems.

Fitts and Posner described the acquisition of psychomotor skills as occurring in three distinct but overlapping stages: the cognitive, the integrative/associative, and the
autonomous phases (Fitts 1967). The cognitive stage involves intellectualizing the task and planning the steps necessary, while the associative stage focuses on practice and repetition, with knowledge of results used to evaluate and guide performance. It was thought that knowledge of results, including the identification and understanding of errors was crucial – without knowledge of results, improvement cannot be obtained (Dunphy and Williamson 2004). Eventually, the learner progresses to the autonomous stage, whereby routine performance requires minimal cognitive input – reaching this stage is skill dependent. It is important to note that this model, while widely accepted, has not been validated in the surgical fields (Dunphy and Williamson 2004).

It is also acknowledged that there is a significant interaction between a learner’s level of interest, and their ability to acquire psychomotor skills, with unmotivated learners unlikely to acquire skills and tending to ‘go through the motions’ (Dunphy and Williamson 2004). Certainly, this puts the focus on instructors to establish the relevance of particular tasks to a learner’s professional career.

In 1990, Miller described a framework for clinical skills assessment (Miller 1990). At the base of a four-tiered pyramid, the learner first ‘knows’, then ‘knows how’. After using that information to ‘show how’, eventually the learner ‘does’ or performs the skill in the clinical setting (ten Cate, Snell et al. 2010). Using Miller’s pyramid, outcomes-based competence is at the top – it is not what the trainee knows, or knows how to do, or even shows, but ‘rather it is a questions of what he or she does’ (Hodges 2010).

2.1.4 CONCLUSION

In this section, varied definitions of what competence can mean in the context of postgraduate medical training have been discussed. It would appear, to this author, that defining competent as a minimum standard, whereby a practitioner has been judged capable of performing a task, is extremely practical in the setting of CBME. Accepting such a definition provides a baseline upon which assess competence in medical residents. Many authors have also interspersed the concept of independence within definitions of competence, suggesting that the ability to perform without supervision is critical. These concepts will form an integral component of this thesis.
The review of models of expertise highlights that the concept of competence is context dependent, and that a gradual transition between stages is a consistent theme. Certainly, learning curves have been identified in the studies of skill acquisition that are available. However, at this time, there is no published literature which describes the development of competency in CBME (Norman, Norcini et al. 2014). This will be an important focus of research for this thesis; can residents achieve a minimal level of competence after a rotation within a CMBE model of training, or is there a more gradual transition to a minimal level of competence that occurs over time, related to experience and duration of training?

2.2 COMPETENCY-BASED MEDICAL EDUCATION

The introduction of CBME at the University of Toronto is one of the first, formal, outcome-based postgraduate medical programs in North America. There are many reasons for this transition from a traditional time-based model, and these potential advantages (and disadvantages) of CBME are discussed in this section. Concepts such as milestones and Entrustable Professional Activities (EPA) are carefully considered, as these concepts may become important in future implementations of CBME. Finally, the current research regarding CBME is summarized, as is the evidence that has emerged from the University of Toronto, highlighting the research deficiencies in this area.

2.2.1 BACKGROUND

The traditional method of orthopedic training, as well as the majority of postgraduate medical training programs in Canada and worldwide, utilizes a time-based system, often combined with a formal exit examination. It is proposed that the advent of CBME will move postgraduate training away from a time-based training model, to a model based upon observable and measurable outcomes (Pimlott 2011).

Competency-based medical education has been defined as an outcomes-based approach to the design, implementation, assessment and evaluation of medical education programs, using an organizational framework of competencies (Frank, Snell et al. 2010). A distinguishing feature of CBME is its dedication to predefined graduate
abilities as the organizing principle (Frank, Mungroo et al. 2010). For this reason, CBME is based upon the successful demonstration and application of specific knowledge, skills, and attitudes that are required for the practice of medicine (Iobst, Sherbino et al. 2010).

This paradigm shift from current time-based curricula to CBME is the Flexnerian revolution of the 21st century (Carraccio, Wolfsthal et al. 2002). The primary goal is to reshape postgraduate training, such that residents become ‘competent by design’, rather than by opportunistic learning (Iobst, Sherbino et al. 2010). ‘Competence by design’ is the title of a series of ten papers that outline a strategic vision of the Royal College of Physicians and Surgeons Canada (RCPSC), a vision to transform postgraduate medical education in Canada to a competency-based model of education and assessment (RCPSC 2014).

While acknowledging the many strengths of current post-graduate training in Canada, the RCPSC offers many reasons for a change to competency-based programs. Most importantly, these documents site an era of increased accountability, with a need to ensure that all graduates are being adequately prepared for practice (RCPSC 2014). In this way, the RCPSC emphasizes the need for curriculum planning to focus on outcomes, rather than fixed time periods for trainees.

Four steps have been highlighted in the development of CBME: 1) competency identification, 2) determination of competency components and performance levels, 3) competency assessment and 4) overall evaluation of the process (Carraccio, Wolfsthal et al. 2002, Leung 2002, Frank, Mungroo et al. 2010). The first step, identification of competencies, can occur through multiple processes, including the Delphi technique, nominal group technique (group consensus), and the simplest method, practitioner surveys (Carraccio, Wolfsthal et al. 2002). Importantly, content development should be performed by faculty experts, and be followed by the creation of a curriculum mapped to an accreditation framework (Iglar, Whitehead et al. 2013). The second step involves determination of competency components, or the tasks that are allocated to each competency (Carraccio, Wolfsthal et al. 2002). These tasks must be measurable, with clearly defined performance criteria and benchmarks for each
competency. The final step involves validating the competencies, and assessment systems (Carraccio, Wolfsthal et al. 2002).

2.2.2 MILESTONES

A systematic review by Frank et al. identified that much of literature discussing CMBE promoted the concept of ‘progression of competence’, whereby learners advance towards competence along a series of defined milestones (Frank, Mungroo et al. 2010). In this way, CBME frameworks require defined levels of progression towards desired outcomes (Ben-David 1999) – the use of milestones is one method that can be used to identify the progression from novice to expert (Lacasse, Theoret et al. 2014).

In the USA, ACGME has been developing the concept of clinical milestones, based upon its six ‘general competencies’ (ACGME 1999, Green and Holmboe 2010). Using the six general competencies as categories of clinical skills that trainees should demonstrate increasing mastery of, milestones are used to dictate the rate at which residents progress through training. These milestones can be used by faculty to determine if a trainee is on an appropriate trajectory, with milestones becoming the blueprint for both the curriculum and the assessment tools (Green, Aagaard et al. 2009).

As part of the Orthopedic Surgery Milestone Project, the ACGME has published a list of key physician competencies for orthopedics (ACGME 2013). In this publication, milestones are presented as knowledge, skills, attitudes, and other attributes for each competency, organized in a developmental framework from less to more advanced. These milestones act as descriptors and targets for resident performance as they move from entry into residency through to graduation, and are numbered from Level 1 to Level 5 (synonymous with moving from novice to expert). The clinical conditions listed are divided into patient care and medical knowledge competencies; separate milestones are used to measure the other competencies of interpersonal and communications skills, practice-based learning, professionalism, and systems-based practice.
As the majority of the research in this thesis will take place in the orthopedic subspecialty of sports medicine, it is interesting to identify the sports medicine conditions described in the Orthopedic Milestones Project; 1) anterior cruciate ligament (ACL), 2) meniscal tear, and 3) rotator cuff injury. The ACL patient care competency describes a Level 1 milestone as a resident that can perform a basic history and examination, and order basic imaging. A Level 2 milestone means that residents can perform a more focused history and examination, as well as appropriately interpret basic imaging. A Level 3 milestone indicates a resident that can recognize concomitant injuries (i.e. lateral collateral injuries), and appropriately order and interpret advanced imaging, and perform diagnostic arthroscopy. A resident deemed to be at Level 4 can perform ACL graft passage and fixation, and is capable of treating postoperative complications, while a Level 5 resident can perform revision ACL and surgically treat complex complications.

It is important to note that milestones do not necessarily correlate with postgraduate year of training, although there is some correlation with experience. For example, Level 1 indicates that a resident is demonstrating the milestones expected of an incoming resident, while Level 3 is described as that of a mid-residency level. Tracking from Level 1 to Level 5 is synonymous with moving from novice to expert (ACGME 2013). Level 4 is described as the goal for graduation, with Level 5 described as an aspirational goal for the few residents who have advanced beyond the performance targets set for residency. Certainly, the Level 5 goal for ACL patient care described above would be far beyond the reach of the majority of residents. While both clinical and technical skills are included in each milestone, it is not clear whether residents must master all components in order to achieve each milestone. Interestingly, it is noted that an overall Level 4 milestone is not described as a graduation requirement – decisions about readiness for practice remain the decision of the program director.

The ACGME milestones project recognizes that learning in residency is a continuum. After the required abilities for CBME in each medical discipline have been organized as competencies, educators can work backwards to identify milestones that trainees will need to reach on the path to acquiring the required competencies (Frank, Snell et al. 2010). However, because some of the procedural skills have inherent dangers, it is
expected that the use of simulator training and assessment will support the determination that a trainee is on the right trajectory (Green, Aagaard et al. 2009, Holmboe, Sherbino et al. 2010).

### 2.2.3 ENTRUSTABLE PROFESSIONAL ACTIVITIES

A critical component of CBME is the requirement for clear lists of critical competencies, following which verification of competency can occur (Long 2000). To bridge the gap between CBME theory and practice, the concept of an Entrustable Professional Activity (EPA) has been introduced (Mulder, Ten Cate et al. 2010).

An EPA describes a professional task that postgraduate residents must master during their training (ten Cate 2005) - tasks or responsibilities that faculty entrust to a trainee to execute, unsupervised, once an adequate level of competence has been achieved (Ten Cate 2014). Typically EPAs are units of work (i.e. manage a patient with an intertrochanteric hip fracture) which describe residents’ abilities to apply either the six multiple competencies described by the ACGME (Goodley 2007, Ten Cate 2014), or competencies within the CanMEDS Roles described the RCPSC (Frank 2005). When put together, EPAs form the mass of critical elements that operationally define a profession (ten Cate and Scheele 2007).

An EPA may be distinguished from a competency by completing this sentence: “Tomorrow the resident will be entrusted to ....” (Mulder, Ten Cate et al. 2010). Once appropriately defined, EPAs may be used to establish a competency-based curriculum, and define five levels of responsibility and proficiency; limited knowledge, act under close supervision, act under supervision on call, act independently, and able to supervise others (ten Cate and Scheele 2007, Sterkenburg, Barach et al. 2010, ten Cate, Snell et al. 2010). The ability to achieve level 4 (acting independently) in predetermined EPAs has been described as a critical goal of CBME (Ten Cate 2014).

Performing an EPA requires a resident to demonstrate a certain level of competency or proficiency in multiple clinical skills to in order to successfully perform an EPA (Chang, Bowen et al. 2013). For example, performing a surgical procedure requires a good knowledge of anatomy, instrumentation, and details of the technical procedure,
as well as organizational skills, an ability to interact with health care professionals, and an ability to communicate professionally with the patient (Mulder, Ten Cate et al. 2010) (Carraccio and Burke 2010). To date, EPAs have been created for pediatrics (Jones, Rosenberg et al. 2011), internal medicine (Chang, Bowen et al. 2013, Lowry, Vansaghi et al. 2013, Caverzagie, Cooney et al. 2015), family medicine (Shaughnessy, Sparks et al. 2013), anesthesiology (Jonker, Hoff et al. 2015), and psychiatry (Boyce, Spratt et al. 2011), and used to guide competency-based assessment (Schultz, Griffiths et al. 2015). As of June 2014, the ACGME requires reporting on selected milestones, with incorporation of EPAs into training programs (Nasca, Philibert et al. 2012, Aylward, Nixon et al. 2014).

At this time, there has been little in the way of published literature examining the use of EPAs in postgraduate residency training. Hauer et al. conducted two pilot EPA-based assessments (inpatient discharge and family meeting) in the clinical setting, testing them on PGY1 internal medicine residents – both the residents and faculty felt it improved skills and facilitated useful feedback (Hauer, Soni et al. 2013). Alyward et al. developed a patient handoff EPA for internal medicine and pediatric interns, identified as a critical skill for residents (Aylward, Nixon et al. 2014). Under direct observation, the interns were judged using the five levels of entrustment as described by ten Cate et al. (ten Cate and Scheele 2007), with the majority judged as being able to perform under direct or indirect supervision. Schultz et al. described the development and use of multiple EPA’s in family medicine, using electronic field notes to structure day-to-day assessment and feedback for residents (Schultz, Griffiths et al. 2015).

The orthopedic faculty at the University of Toronto has created a list of the top ten EPAs for the program, which all residents are required to be able to perform unsupervised, prior to graduation from the program. How to assess residents’ ability to perform these EPAs independently is unknown. One difficulty is that many EPAs are longitudinal and describe care that unfolds over time (Chang, Bowen et al. 2013). In the clinical setting, components of many EPAs would require assessment in different settings and at different times, posing logistical difficulties. For example, a resident might assess a patient for a total knee arthroplasty (TKA) preoperatively, and perform the technical procedure on a different patient later that week. The use of
simulation to create and assess competence in an entire EPA is one potential method to overcome these assessment difficulties, allowing faculty to determine a resident’s ability to perform tasks prior to performance in the clinical setting, as well as determine the appropriate level of supervision required.

2.2.4 PROPOSED ADVANTAGES OF CBME

Many postgraduate training programs are moving towards a CBME model (Cunningham, Kates et al. 2014), for a variety of potential benefits. These benefits include improved training and resident performance, the ability to identify outlying residents requiring remediation (Lacasse, Theoret et al. 2014), as well as the potential to shorten training. While CBME appears set to become the standard in postgraduate training in Canada, there is a lack of evidence with respect to any significant outcomes (Iglar, Whitehead et al. 2013).

Curriculum planning

A major advantage of CBME is that it ensures that curriculum planning is explicitly tied to the outcome needs of graduates – educators can then use this competency framework as a guide for designing learning experiences (Frank, Snell et al. 2010). Residency programs classically have curriculum guidelines, without clear definitions of the expected competencies (Carraccio, Wolfsthal et al. 2002). Once relevant competencies have been defined for each specialty, educators can work to ensure training and assessment is aligned with these predefined outcomes. In this way, a carefully designed curriculum can ensure that residents are, at the very least, minimally exposed to all essential components of training (Norman, Norcini et al. 2014).

Reduced focus on exit exams

Traditionally, a resident’s competency has been measured by the program director’s opinion that the resident is ready to practice independently (Long 2000). This typically occurs after a specified number of years in training, supported by successful completion of a specialty board exam (Long 2000). Moving to a CBME model can help to ensure that medical graduates have demonstrated competence in the critical
domains of their subspecialties prior to certification examinations (Frank, Snell et al. 2010). Certainly, exit examinations are limited in both time and scope, allowing only a sampling of some of the critical components of training. Furthermore, at this time, exit examinations for most surgical programs do not assess competency in the performance of technical procedures, an obvious concern. For this reason, regular assessment throughout training can complement exit examinations, and potentially, in the future, eliminate the need for these costly assessments.

**Accountability**

Another advantage of CBME is transparency, whereby the medical profession and the public can be more confident that training programs are producing competent physicians (Norman, Norcini et al. 2014). This theme was identified in a systematic review by Frank et al., whereby authors identified that the CBME model is the right choice in an ‘era of greater accountability to the public’, with this training model organized around competencies needed to serve patients (Frank, Mungroo et al. 2010). Certainly, it is believed by some that the slow ascent to competence through trial and error is no longer acceptable – residents must demonstrate competence in critical tasks of their specialty prior to transition to unsupervised practice (Hall, Ellis et al. 2003, Schaverien 2010).

**Focus on the learner**

The concept of CBME has also become increasingly popular among educators due to its focus on the learner (Leung 2002, Frank, Mungroo et al. 2010, Frank, Snell et al. 2010). Competency-based medical education demands an increase in instructional methods, meaningful supervision by expert faculty, and increased engagement of both the faculty and the resident (Iobst, Sherbino et al. 2010). Regular assessment also provides both formative and summative feedback, allowing residents and faculty to reflect on performance. In this manner, regular assessment in the setting of CBME is critical, and used to identify deficiencies in resident performance, as well as educational deficiencies in a training program.
The resident in difficulty

In the traditional time-based programs, trainee problems are often not recognized until the later stages of training; regular assessment of competency can allow the identification of deficiencies much earlier, to the benefit of both the trainee and the program (Holmboe, Sherbino et al. 2010). In a study of the Society of Neurological surgeons, the outcome of training was examined using exit surveys of trainees and program directions; in the opinion of the program directors a small number of trainees possessed questionable competence, while up to 10% of trainees questioned their own competence to perform procedures (Bosc 1999). In another study surveying 91 subspeciality program directors in North America, it was felt that 30% of new fellows could not perform a laparoscopic cholecystectomy independently, and that 66% were not able to operate in a major procedure unsupervised for 30 minutes (Mattar, Alseidi et al. 2013). These papers support the notion that time-based training programs are not adequately serving a percentage of trainees. In this manner, regular assessment in the setting of CBME, and identification of the resident in difficulty, is a major potential advantage.

The CBME concept demands that all residents remain in training until he or she has been shown to have the required knowledge and skills, with the ability to apply them independently (Long 2000). In the (rare) instance of trainees are not able to obtain a minimal level of competence, regular assessment can allow identification early in training, facilitating a fair exit from a training program (Holmboe, Sherbino et al. 2010). In the setting of residents leaving training programs, it is also important that assessments used to support such an end-point are objective, defensible, and carried out by multiple observers over multiple time points.

Time

Another proposed advantage of CBME is the potential to shorten the length of medical residencies to the time taken for a resident to complete their checklist of various competencies – a proposal that has both supporters (ten Cate and Scheele 2007) and opponents (Brooks 2009). Certainly, one of the key features of CBME is the ability to allow learners to progress at their own rate (Iobst, Sherbino et al. 2010) - some
residents can achieve competency status earlier than required, while others require increased educational time (Ebert and Fox 2014). Reduced time may only apply to a few residents, in whom frequent, summative assessment will allow rapid advancement (Holmboe, Sherbino et al. 2010). It also believed that training can be shortened by removing non-essential rotations from curricula (Grantcharov and Reznick 2009).

This is an important issue to both educators and administrators. It has been estimated that a year of surgical training costs $100,000 per year per resident – thus any increased costs of CBME could possibly be counteracted by residents completing training in reduced time (Ferguson, Kraemer et al. 2013). However, some authors speculate that the net effect of CBME may be to lengthen training (Taber, Frank et al. 2010). At this time, it may be premature to predict the effect of CBME on time in training, prior to the introduction of valid assessments of competence.

2.2.5 POTENTIAL DISADVANTAGES AND PROBLEMS

Competence v Proficiency and Expertise

There has been concern that CBME will create a culture of emphasizing competency rather than expertise (Cooper 2006). For example, defining competency as achieving a minimum standard in the task being assessed does not necessarily satisfy public expectations (Brooks 2009), in the same way that defining satisfactory as ‘adequate’ is inferior to ‘satisfying expectations and leaving no room for complaint’; clearly in CBME the second is preferred (Crebbin 2005). It is commonly believed that the achievement of competence in predefined areas is the only first step on the road to professional excellence – an important concept that both educator and learner must be aware of (Dreyfus 1986, Talbot 2004, Frank, Snell et al. 2010).

Reductionism

There is a great potential for CBME to overcomplicate medical training. Dividing competencies into key competencies, enabling competencies, and milestones risks creating an exhaustive level of detail that may prove overwhelming and impractical (Frank, Snell et al. 2010, ten Cate, Snell et al. 2010, Lurie 2012). Furthermore, this ‘reductionism’, or reducing clinical training into smaller and smaller components, may
mean that it will never be feasible to conduct enough observations to “sign off” on all competencies (Frank, Snell et al. 2010, Norman, Norcini et al. 2014). For these reasons, one clear risk of CBME is that the unique relationship between master clinicians and apprentices might be replaced by overly simplistic checklists applied to an idea of measurable competencies (Lurie 2012).

**Difficulties faced in the assessment of competence**

There are many inherent difficulties that will be faced when designing assessment tools for the competency-based era. While one of the appeals of CMBE is the development of consistent standards, it is important that educators resist placing an emphasis upon those criteria or tasks that are most easily measured (Talbot 2004) - limiting focus on competencies that can be measured, rather than those that must be learned (Carraccio, Wolfsthal et al. 2002). Examples include qualities such as altruism, humanism, professionalism and scholarship, which are difficult to define and objectively measure, and thus risk limited attention (Norman, Norcini et al. 2014). For this reason, while the seven CanMEDS Roles and six ACGME competencies may appear simple, they are actually rather complex (Norman, Norcini et al. 2014). Competencies are a negotiation of language based upon expert opinion, experience and consensus, making it difficult to develop valid and reliable assessments (Lurie 2012). These issues will be addressed in this thesis.

**Time**

The idea that residents can readily become competent in short focused periods has been questioned. Norman et al. uses a child development analogy, whereby "a child does not crawl one day and walk the next: walking is a gradual process of mastery that occurs over weeks or months" (Norman, Norcini et al. 2014). However, in the CBME program at the University of Toronto, an orthopedic resident must remain in an arthroplasty rotation until they can successfully complete a total hip replacement (Hodges 2010) – a skill that usually requires fellowship training, and takes years to perfect. As Hodges writes, the degree of flexibility required to accommodate all learners achieving competence in all rotations may in fact lengthen training in many instances.
For this reason, an important problem faced by adaptation of a true competency-based program is one of practicality and logistical chaos (Frank, Snell et al. 2010). A purist model of CBME allows learners to progress at their own rate, with some residents advancing early, with more time provided for challenged learners (Iobst, Sherbino et al. 2010). However the fixed and rigid structure of most time-based programs make the adoption of a pure CBME program problematic (Iobst, Sherbino et al. 2010). Many residency programs depend upon a time-based approach in order to provide service and manage call schedules (Taber, Frank et al. 2010). Furthermore, alternative funding models, which are traditionally based upon set periods of time, may need to be explored (Taber, Frank et al. 2010).

For this reason, a hybrid model has been recommended, whereby minimum and maximum time periods for progression are retained. In the series of white papers, ‘Competency by Design’ issued by the RCPSC, it is proposed that CBME at the residency level ‘be incorporated within a time-based structure’, thus advocating for a ‘hybrid model of competency-based, timed rotations’ (RCPSC 2014). In this model, timed rotations remain as a resource for acquiring desired competencies, with summative assessment used to ensure the attainment of competencies and required milestones. Certainly, this would seem an approach that incorporates the advantages of CBME with the advantages of the time-based approach.

**Faculty Development**

Finally, CBME will require greater faculty involvement, in terms of direct observation and assessment. The implication is that clinical teachers will need to spend more time teaching, taking time away from already underserviced enterprises (Taber, Frank et al. 2010). Adoption of CBME programs will also necessitate investment in faculty development, to allow them to understand, implement and facilitate a competency-based approach (Taber, Frank et al. 2010). How this will be funded, and what the implication is on patient care, is currently unknown.
2.2.6 SUMMARY OF THE EVIDENCE REGARDING CMBE TO DATE

While the widespread implementation of CBME mandates that educators demonstrate advantages of CBME, at this time, there is minimal evidence to demonstrate improved outcomes from CBME (Greaves and Loquist 1983, Thurman and Sanders 1987, Smith and Fuller 1996, Iglar, Whitehead et al. 2013). CBME, however, is not a new concept. As early as 1987, Thurman and Sanders compared a group of five radiology technicians receiving a CBME instructional method to six technicians receiving traditional instruction methods – on a post-test assessment, the CBME group had higher scores (Thurman and Sanders 1987). In 1996, the Brown University School of Medicine adopted a CBME training program, defining nine abilities that all students had to attain prior to graduation (Smith 1999). These were translated into observable behaviors, rated at three levels of competency – beginner, intermediate, and advanced (Smith, Dollase et al. 2003). Using results on USMLE as the outcome measure, students were at the national mean, and their pass rates slightly above the mean.

Martin et al. studied the effect of using a competency-based instructional module on the acquisition of three invasive procedures (endotracheal tube insertion, venous cutdown, and insertion of a chest tube) in a general surgery residency program (Martin, Vashisht et al. 1998). Using group instruction and hands on teaching in a cadaver laboratory, this study demonstrated a reduction in failure (inability to perform procedure either correctly or in less than 120 seconds) and complication rates in all three procedures performed in the laboratory, as well as a reduction in the residents’ trauma resuscitation time. In 2013, Singer et al. published a study of critical care residents - following an intensive, three month, simulation-based education intervention, first-year residents outperformed traditionally trained third-year residents on an end of rotation clinical skills assessment (Singer, Corbridge et al. 2013).

Long reported on the experience of the Department of Neurosurgery at John Hopkins, comparing competency-based training and traditional training for the acquisition of procedural skills in neurosurgery (Long 2000). A blueprint was created for a list of five critical procedures, with arbitrary six-month blocks maintained. In the traditional training, the exposure and closure components of procedures were performed during
PGY-2; complete procedures were allowed during PGY-3 and PGY-4. In the competency-based training, progression in responsibility in the OR was allowed when the program director (and the trainee) were confident that the trainee was competent (able to carry out the procedure independently, with supervision). It was reported that by using the competency-based training model, the time to perform complete procedures was reduced by 3 to 6 months for exposures and closure, and reduced by 6 to 9 months for the performance of complete procedures.

While these authors are to be commended on their innovative studies, clearly there are many areas that require research in the field of CBME. The limited amount of literature to date has focused on the assessment of a few technical procedures, and often describes the attainment of competence without the use of valid assessment tools. At this time, there is no published research that examines the ability of residents to achieve a minimal level of competence after training in a CBME program.

2.2.7 THE TORONTO EXPERIENCE

Beginning in 2009, the University of Toronto division of orthopedics began a trial of an outcomes-based residents training program. Some of the impetus for this change to a CBME model came from a reduction in working hours for residency training, with many authors believing that a reduction in working hours would mandate lengthening of residency training programs (Pellegrini 2012). The division of orthopedics believed that a more appropriate response would be to maximize the educational use of the available training hours - for this reason, a program that allowed residents to move from rotation to rotation based upon demonstrated competency rather than after a fixed time period was implemented (Alman, Ferguson et al. 2013).

The so-called competency-based curriculum (CBC) was based upon three main principles: 1) teach the appropriate spectrum of competencies required, 2) maximize the use of available training time, and 3) allow residents to progress through the program at their own rate (Alman, Ferguson et al. 2013). The aims of the new program were to make training modular, to accelerate the speed of technical skill acquisition, to reduce inefficiencies in training (primarily by removing off service or non
orthopedic rotations), and to provide frequent formative and summative assessment (Ferguson, Kraemer et al. 2013).

Early in the CBC program, the importance of clear definitions was recognized. In an instructional course lecture written by the orthopedic division, competency was defined as having the necessary ability, knowledge, and/or skill to manage a clinical situation, but not necessarily in an independent manner (Alman, Ferguson et al. 2013). Proficiency was defined as the demonstrated ability to undertake a task in an independent manner (clinical fellow), while expertise was defined by extensive knowledge or ability beyond what is expected from the average clinician (and expected of the orthopedic faculty. It is important to note that the origins of these definitions were not referenced in this article. Alman et al. also stated that “the purpose of a residency program is not to create clinicians who are proficient at all subspecialty procedures, but …… proficient at skills (such as treating hip fractures) that a general orthopedic surgeon would be expected to perform in his or her first months in practice” (Alman, Ferguson et al. 2013). It was also stated that “the program was structured so that residents could advance through training as quickly as possible” (Alman, Ferguson et al. 2013).

Using the core curriculum of the RCPSC as a baseline, the curriculum was divided into 21 modules, in three phases – trainees had to complete phase one modules before starting phase two modules (Alman, Ferguson et al. 2013) (Appendix A). A multidimensional assessment algorithm was used to distribute assessment methods throughout the modules, using a combination of multi-choice questions (MCQs), structured oral examinations (SOEs), Objective Structured Assessments of Technical Skills (OSATS), multisource 360-degree feedback, patient assessment, and observed history and physical examinations.

The purpose of these assessments was described as two-fold. Firstly formative, and used to provide the opportunity for guided and focused self-improvement. Secondly summative, to confirm that specific benchmarks of performance had been obtained (Ferguson, Kraemer et al. 2013). While the benchmarks used were not described, Ferguson et al. stated that “residents were administered RCPSC-level oral
examinations pertinent to their current module.... designed to assess application of acquired knowledge in clinical situations and managing complications" (Ferguson, Kraemer et al. 2013). Ferguson et al. emphasized that the demonstration of competency was the key underlying requirement for advancement (Ferguson, Kraemer et al. 2013).

In a pilot study beginning in July 2009, 3 of 12 incoming residents were trained under the new CBC program – this occurred each year for three years (Alman, Ferguson et al. 2013). As of 2013, 14 residents were enrolled in the CBC program (Ferguson, Kraemer et al. 2013). Of the first three residents, two completed all 21 modules in four years, compared to the RCPSC mandated five years – these residents were given permission to sit, and consequently passed their fellowship exams early (Ferguson, Kraemer et al. 2013). Just as importantly, the intense assessment process identified residents with deficiencies early in the program – as a result two of the 14 residents took longer than expected to complete three different modules, largely due to knowledge deficits (Ferguson, Kraemer et al. 2013).

There is little in the way of objective evidence to support the new program, with these review articles reporting largely anecdotal results. It has been stated that there is an extremely high degree of satisfaction with the new model of training, with traditional stream residents wishing to participate in the educational programs of the CBC, due to its perceived advantages (Alman, Ferguson et al. 2013). Reportedly, faculty generally believed that trainees in the CBC program were better prepared to perform surgery in the OR. It was also reported, again without objective evidence, that CBC residents had treated between two and three times the number of hip fracture patients (Alman, Ferguson et al. 2013). Ferguson et al. also wrote that there was a trend towards better performance by CBC residents on the Orthopedic In-Training Examination (OITE) results, an international, orthopedic MCQ undertaken every year by all orthopedic trainees (Ferguson, Kraemer et al. 2013).

A series of publications from the division of orthopedics at the University of Toronto examined the use of an OSATS to assess all incoming orthopedic residents' ability to perform casting, drape limbs, expose soft tissues, use power tools, close wounds, and
communicate in the operating room (Sonnadara, Van Vliet et al. 2011). Following this OSATS, the residents enrolled in the CBC pilot program were exposed to an intensive surgical skills course – these residents (n=6) subsequently outperformed the non-CBC residents (n=6) on an OSATS reassessing core surgical skills. A follow-up study examining retention of these skills found that, seven months later, the CBC group continued to outperform the non-CBC group, with no difference between the CBC group and a group of senior residents (Sonnadara, Garbedian et al. 2012). While these studies serve to highlight that an intensive skills course is highly effective at teaching and developing targeted technical skills in first-year residents, it does not demonstrate that CBME is superior to traditional training programs.

It is acknowledged that CBME at University of Toronto is in its initial phases, with insufficient time or outcome measures available to compare the new program to the traditional program – the first step is to determine that residents in the new program being trained to an equivalent standard as those within the traditional program (Alman, Ferguson et al. 2013). There continue to be many unresolved issues faced by the new program, similar to those discussed in the CBME disadvantages section of this literature review – these include increased costs, the burden of regular assessment, and the need for increased faculty involvement.

2.2.8 CONCLUSIONS

Despite the many potential advantages of CMBE, there are clear challenges faced by the change from a time-based system to an outcomes-based one. The first that is faced by the division of orthopedics at the University of Toronto, and all other programs adapting CBME, is the development of valid assessment tools that focus on predefined critical competencies - only the development of these assessments will allow objective judgments regarding resident competency be made. The second challenge will be the maintenance of an outcomes-based program that works logistically, while continuing to ensure a focus on the learner. One of the major advantages of a true CBME model is that each learner can take as much time as required in order to reach competency. Whether competency is within the reach of all residents is unknown.
The sports medicine rotation offers an opportunity to develop and evaluate assessment tools within a CBME setting. Residents at the University of Toronto undertake two three-month sports rotations during their training, once as a junior (postgraduate year 1 – 3) and once as a senior (postgraduate year 4 – 5). Residents are expected to demonstrate a minimal level of competence at the end of their rotation, covering a wide range of competencies. Critically, issues such as validity, feasibility and acceptability must be addressed.

2.3 PSYCHOMETRIC & ASSESSMENT THEORY

A deficiency in the CBME literature to date is the limited use of objective assessments to determine a minimal level of competence in residents. Van der Vleuten introduced a conceptual model defining the utility of assessment instruments, describing five important characteristics; validity, reliability, educational impact, cost and acceptability (van der Vleuten 1996). This section will discuss the various components of psychometric theory as they pertain to designing modern assessment programs in the setting of CMBE. Importantly, it will compare and contrast classical and contemporary concepts of validity, important when evaluating a new assessment program. Furthermore, this section will discuss issues such as educational impact, cost and acceptability, which are important aspects of the sustainability of a CBME assessment program.

2.3.1 VALIDITY

The validity of an evaluation scheme is the extent to which the methodology actually measures what it is supposed to measure (Bhatti and Cummings 2007), or the degree to which a result reflects the construct it is supposed to measure (Crossley, Humphris et al. 2002). Validity may also refer to the degree to which the conclusions derived from an assessment are well grounded, justifiable, relevant and meaningful – how well can results of a test as interpreted for a specific purpose be trusted? (Cook, Beckman et al. 2010). Without validity, assessments in medical education have no meaning (Downing, Lieska et al. 2003).
Validity is not a property of an instrument, but rather of an instrument’s scores and its interpretations – thus validity must be sought for each intended use or interpretation (Cook, Beckman et al. 2010). In order to determine if the use of an instrument is valid, evidence must be presented to support or refute the meaning or interpretation assigned to assessment results (Downing 2003).

Classical concept of validity

Previously, validity was broken down into components. For example, face validity was considered when an assessment tool looked like it might measure what it was intended to (Crossley, Humphris et al. 2002) - in this way, face validity was used to describe the appearance of validity in the absence of testing (Cook, Beckman et al. 2010). Content validity was demonstrated by a systematic approach to ensuring an instrument assessed a particular construct, in association with relevant performance criteria (Crossley, Humphris et al. 2002, Downing and Haladyna 2004). Criterion validity was established by comparing the assessment with several other measures that tested similar elements of performance (Crossley, Humphris et al. 2002). Criterion validity was subdivided into concurrent (correlates well with a measure that has previously been validated) and predictive (used to predict a later measure) depending on the timing of data collection (Downing 2003).

Construct validity was associated by demonstrating a reasonable hypothesis i.e. that surgical skill improves with experience and year of training (Crossley, Humphris et al. 2002). Lurie et al. believe the finding that more experienced test takers achieve higher scores was the weakest possible evidence about whether results represent a particular hypothesis or competency (Lurie, Mooney et al. 2011). However, Green and Holmboe have endorsed assessment methods that are able to discriminate between different levels of performance within a single competency, on the basis that such assessments are useful when charting residents’ progression through different milestones of competence (Green and Holmboe 2010).
Contemporary concepts of validity

The modern theory of test validity approaches validity as a hypothesis, with data used to support or refute the validity hypothesis, making an assessment more or less valid (Downing 2003). In this way, validity becomes an overall concept, whereby validity relates to the number of logical and statistical arguments that can be made to support a particular interpretation of test results (Lurie, Mooney et al. 2011). In this way, validity of an instrument is not dichotomous - validity is always a matter of degree, as an instrument will never be able to perfectly reflect the underlying construct. Thus, an assessment is never ‘valid’ or ‘invalid’; evidence is used to support or not support the proposed interpretation of the scores (Downing 2003). Validity can never be proven (Cook, Beckman et al. 2010).

In a contemporary concept of validity, ‘the process of validation is closely aligned with the scientific method of theory development, hypothesis generation, data collection for the purpose of hypothesis testing and forming conclusions concerning the accuracy of the desired score interpretations’ (Downing 2003). Multiple sources of evidence are required, and the amount of evidence necessary varies according to the intended purpose of the instrument – for example, more evidence of validity is required in high stakes exams (Downing 2003, Cook, Beckman et al. 2010). In the modern framework, all validity is construct validity, because assessment in medical education deals with constructs, defined as intangible collections of abstract concepts and principles inferred from behavior and explained by educational and / or psychological theory (Downing 2003).

Thus the framework of validity has moved from types of validity (face, content, and criterion) to a concept of identifying evidence for the validity of results and their use (American Educational Research Association, Association et al. 1999, Korndorffer, Kasten et al. 2010). This unitary concept of “construct validity” (Messick 1989) uses evidence to support a validity argument collected from five sources (Downing, Lieska et al. 2003, Cook, Beckman et al. 2010) (Messick 1989, American Educational Research Association, Association et al. 1999):
1. **Content**: do instrument items completely represent the construct? In education assessment, this involves using a detailed test blueprint, ensuring the test is representative of the domains, using qualified writers to formulate well-written questions, and directly linking item content to the curriculum. In the setting of an OSCE for example, the cases should be reviewed by independent content experts, ensuring that the cases represent both the blueprint and the intended construct. Ultimately, these steps ensure the quality of the test questions. Content evidence is often presented by providing a detailed description of these steps.

2. **Response process**: the relationship between the intended construct and the thought processes of subjects or observers. The response process should ensure accuracy of the responses to the assessment, with quality control of the data. This involves student familiarity with the exam format and clarity of instructions - asking students to ‘think aloud’ may be an important process. Other aspects include test security, rater training, clear rater thought processes, quality control of scoring, electronic scanning, and score reporting with accurate descriptions and interpretations of scores. Documenting the rationale for using scoring methods (i.e. global ratings v checklists) is important.

3. **Internal structure**: relates to the psychometric properties of an assessment, such as acceptable reliability and factor structure. Without high reliability, or reproducible scores, validity evidence is lacking. Internal structure may be demonstrated by item analysis (item difficulty, item-discrimination, inter-item correlations), internal consistency, inter-rater reliability, standard errors of measurement (SEM), and generalizability theory. Generalizability theory is very useful in performance examinations, as it can be used to estimate the contribution of error from raters, cases, items or examinees.

4. **Relations to other variables**: correlation with scores from another instrument assessing the same construct. Typically, this involves correlation with scores on other performance assessments, such as an existing older measure with accepted characteristics. These may be convergent correlations (similar tests) or divergent correlations (dissimilar measures).
5. **Consequences**: do scores really make a difference, is there an important decision to make, or will the assessment impact teaching and learning? Some licensing examinations have very significant consequences i.e. failure to pass and the need to repeat training and examinations. Assessment tools must be consequences, with positive consequences outweighing the negative ones, which is of course subjective. All assessment must therefore describe the method of determining pass/fail scores as well as the rationale for choosing any particular method. The passing score should be acceptable (again subjective), reliable, reasonable, and defendable.

Validity is important when designing and evaluating assessment tools, whether in the setting of an OSCE or in the use of surgical simulation (Korndorffer, Kasten et al. 2010). However, in 2010 a review of the literature with regards simulators in laparoscopic surgery identified that only 23% adhered in part to the ‘new’ consensus standards for validity, with the majority continuing to use old frameworks including construct, face, and content validity (Korndorffer, Kasten et al. 2010).

**Threats to validity** may be divided into construct under-representation (CU) (inadequate sampling of the construct domain) and construct irrelevant variance (CIV) (bias exerting nonrandom influence) (Messick 1989, Downing and Haladyna 2004). In performance-based assessment, examples of CU include too few cases or unrepresentative cases, while examples of CIV include flawed checklists, inappropriate case difficulty, poorly trained raters, rater bias, rater error (halo, severity, leniency, and central tendency error (ratings in the centre of the scale, restriction of range) (Downing and Haladyna 2004). These issues are discussed further in the section below on reliability.

**2.3.2 RELIABILITY**

Reliability is defined as the reproducibility or consistency of scores from one assessment to the next (Cook, Beckman et al. 2010). Like all scientific methods, assessments must be reproducible in order to be interpreted correctly or meaningfully - (Downing 2004). While the contemporary concepts of validity include reliability, as discussed above, this thesis will discuss issues of reliability separately, because
issues of rater bias and inter-rater reliability are important issues in a CBME assessment program.

Reliability is a prerequisite of, but insufficient for, validity (Downing 2003, Cook, Beckman et al. 2010). Reliability can be measured in many ways, depending on the instrument type. There are many types of reliability, including test-retest, split-half, internal consistency, as well as inter- and intra-observer reliability.

Theoretically, reliability is defined as the ratio of the true score variance to the total score variance – reliability coefficients are thus used to estimate the amount of measurement error in assessments (Downing 2004). Reliability is usually reported as a coefficient ranging from 0 (no correlation, all error) to 1 (perfect correlation, with all variance related to the subjects undertaking the instrument) (Traub 1991). For very high stakes assessment (such as licensing and certification), reliability should be greater than 0.9; for moderate stakes assessment (such as end of year medical school exams) reliability should be greater than 0.80 (Nunnally 1994). For less important settings (such as end of rotations OSCE’s), reliability should be 0.7 or more (Downing 2004). It is important to note that these are guidelines only.

Measures to improve low reliability involve increasing the number of items, questions, observations, cases or raters (Downing 2004). Reliability is also increased by using clearly written questions, content expert review, and by using questions of medium difficulty – questions that are either very easy or very hard (when students get most correct or incorrect) tend to generate low reliability. The expected improvement of adding items can be predicted using Spearman-Brown formula (Traub 1991). Another factor known to increase reliability is having a heterogeneous sample of examinees with a range of performances. An example is known-group comparisons (expert-novice comparisons) which can serve to increase the reliability coefficient (Cook 2015).
**Test-retest**

Test-retest is commonly used to estimate internal consistency, but is generally inappropriate in education measurement, as it is assume that most students will learn after the first test (Downing 2004).

**Internal consistency**

Internal consistency is a measure of how well all items on an instrument measure the same construct (Tavakol 2011). Internal consistency is a form of split-half reliability, whereby the reliability can be calculated after a single test or single application of an instrument, by seeking correlation between two random half tests (i.e. the first and second halves of an assessment), and using them as a proxy for two complete tests given to the same group of participants (Downing 2004, Cook, Beckman et al. 2010).

Cronbach’s alpha is an internal consistency estimate of the reliability of test scores, or how closely related sets of items are as a group (Downing 2004). Cronbach’s alpha estimates the internal consistency by using all the possible ways to split a test into two halves, and taking the average of the correlations between scores on each half (Downing 2004). In this way, internal consistency becomes one way of finding all the possible split halves. This interpretation of reliability is the correlation of the test with itself – squaring this correlation and subtracting from 1.00 gives an index of measurement error. As the reliability increases, the percentage of a test score that is a result of error decreases (Tavakol 2011). Typically, scores measuring a single construct correlate highly, conversely, a low internal consistency may mean that more than one construct is being measured (Cook, Beckman et al. 2010). In the setting of a test that has more than one concept or construct, either alpha should not be calculated, or it should be calculated for each separate concept rather than the entire test (Tavakol 2011).

Alpha may not always be the most appropriate estimate of reliability in OSCEs because of two reasons; 1) it examines only a single facet of measurement error, and 2) it considers the rank order of examinees rather than their deviations from a standard
(Brannick, Erol-Korkmaz et al. 2011). However, it remains the most commonly reported index of reliability.

**Generalizability Theory**

The most elegant estimate of inter-rater agreement uses generalizability (G) theory, which can be used to estimate the variance components for all the variables of interest (Downing 2004). Generalizability theory examines what proportion of the error in the measurement is a result of various factors involved in the measurement process - there are four major sources of variance: items (individual questions on an instrument), forms/scales, subjects, and raters (Downing 2005). In other words, G theory uses analysis of variance to quantify the contribution of each error source to the overall error or unreliability of the assessment. In this way, G theory can quantify the relative size of task-irrelevant influences on examinee’s scores, allow estimation of multiple sources of error, and determine the impact of internal consistency and inter-rater reliability on an assessment tool (Cook, Beckman et al. 2010).

Generalizability theory can also be used to provide an estimate of reliability that takes all factors into account simultaneously, and to calculate the inter-rater reliability (and subsequent rater variance) after eliminating other known sources of measurement error (Cook, Beckman et al. 2010). The results of a G-study can also be used to run a decision analysis, whereby the impact on reliability if some of the factors were manipulated can be estimated – in this way G theory can be used to estimate the resources required i.e. number of stations or increased examiners per stations.

**Inter-observer reliability**

All performance evaluation depends upon the consistency of raters and their ratings for reliability – for this reason, the largest threat to reproducibility is low inter-rater consistency (Downing 2004). The simplest measure of inter-observer reliability is percent agreement, but this calculation does not account for agreement by chance. Alternatively, the Kappa statistic (a type of correlation coefficient) can be used, which can account for random chance agreement (Downing 2004). The intra-class correlation coefficient (ICC), which uses analysis of variance (ANOVA) to estimate the
variance associated with factors in the reliability design, is more commonly used as it can estimate the inter-rater reliability of multiple raters, as well as the reliability of a single rater (Downing 2004, Cook, Beckman et al. 2010).

**Rater error / Examiner Bias**

Objectivity is a generic goal of assessment or measurement, marked by freedom from subjective influences - objectification describes strategies used to reduce measurement error (Norman, Van der Vleuten et al. 1991). True objectivity is very difficult to obtain in most assessments. The assessment of competency always requires the exercise of judgment by another competent, proficient or expert practitioner; for this reason the development of an objective assessment, or assessment of competency independent of the assessor is not always possible (Brooks 2009). An attempt to minimize these subjective effects is made via a process of standardization – by reaching consent with regards to what is being evaluated and the standards expected (Ginsburg, McIlroy et al. 2010).

It is thought that doctors who judge trainees largely agree on performance, but only when judges have appropriate experience in assessment (Crossley, Humphris et al. 2002). Furthermore, it is important to note that score variation may be viewed as a form of idiosyncrasy rather than error (Govaerts, van der Vleuten et al. 2007), caused by multiple individual traits of observers (Yeates, O'Neill et al. 2013), or multiple traits of different residents (Ginsburg, McIlroy et al. 2010). However, of all the sources of variance in rating assessments, variance due to raters is typically the largest component, accounting for 80-90% (Govaerts, van der Vleuten et al. 2002, Downing 2005). For example, one study by Kreiter and Ferguson determined that that an average of 89% of the total variance was attributed to the raters, with only 10% due to the subjects (Kreiter and Ferguson 2001).

There are many examples of rater error or bias including halo effect, examiner leniency or stringency (doves & hawks) (McManus, Thompson et al. 2006, Harasym, Woloschuk et al. 2008), central tendency error (ratings in the centre of the scale), and restriction of range) (Downing and Haladyna 2004). Each of these sources of error may affect a particular assessment tool.
The *halo effect* refers to when a resident with a good faculty relationship or a strong work ethic receives higher ratings in unrelated fields (Sidhu, Grober et al. 2004). It has been shown that a positive relationship of examiners with candidates can be associated with significant increase in ratings, up to a difference between a first and third/fourth year resident (Stroud, Herold et al. 2011). For this reason, most high-stakes examinations avoid pairing examiners with familiar candidates, however this can be impossible in low-stakes examinations due to examiner availability and scheduling difficulties (Stroud, Herold et al. 2011). The halo effect (and *reverse-halo effect*) has also been demonstrated to occur in a study comparing station-specific versus candidate-specific examiners (Touchie, Humphrey-Murto et al. 2010).

Another common category of rater error is the *hawks & doves effect*, whereby some examiners are particularly stringent, while others are more lenient (McManus, Thompson et al. 2006). In a study of 10,145 residents undertaking physician exams with two examiners marking each station, researchers estimated that 12% of the main effect variance was due to differences in examiner leniency-stringency – in this study greater stringency was associated with greater examiner experience (McManus, Thompson et al. 2006). Importantly, it was noted that the overall effect of this rater error was limited in to regards to the pass-fail rates of candidates, due to an averaging over multiple stations. In a separate study which videotaped general practice trainees, Campbell and Murray noted that while there was evidence of a hawks and doves effect, again there was a high level of agreement with regards to pass / fail (Campbell and Murray 1996).

Competency-based medical education requires that judgements be made against a fixed level of ability; however, there are other forms of bias that may affect these judgements. One such form is *anchor bias*, whereby a recent experience causes the examiner to look for similar actions in subsequent performances (Chapman and Johnson 1999). A similar form of potential bias is *contrast bias*, whereby raters make assessments using comparison with recent performances (Yeates, O'Neill et al. 2013). For example, Yeates et al. demonstrated that raters exposed to videos of good performances were subsequently more likely to rate borderline performances lower (Yeates, O'Neill et al. 2012). This effect may be important, as candidates will often
follow good or poor candidates sequentially through stations in the setting of an OSCE (Newble 2004).

Means of overcoming issues of bias include rater training, increased rater experience, pairing examiners with unfamiliar candidates, having two examiners, or using video-and/or audiotaping (Stroud, Herold et al. 2011). With regards to rater training, two randomized trials have examined this effect, with one demonstrating no significant effect (Cook, Dupras et al. 2009), and the other showing an inconsistent increase in rater stringency (Holmboe, Hawkins et al. 2004). However, there is some evidence that rater training can reduce errors related to range restriction errors (Raj and Thorn 2014).

While having two examiners at a station is potentially advantageous, averaging two or more independent raters may only work to cancel systematic errors (Downing 2005). However, it has been shown that group assessment improves inter-rater reliability in comparison to individual assessments (Thomas, Beckman et al. 2011). Research also suggests that conducting assessments for formative rather than summative purposes reduces leniency (Jawahar 1997), while giving feedback face to face rather than in writing induces leniency (Klimoski 1990).

2.3.3 FEASIBILITY

Feasibility is defined by the Collins English dictionary as the state or degree of being easily or conveniently done; feasible may be used to mean that a task or process is practical, possible, reasonable, viable, achievable or attainable. In the business world, a feasibility study is an analysis and evaluation of a proposed project to determine if it is both technically feasible, and feasible with regards to cost.

The discussion of feasibility in the assessment literature generally revolves around costs, and other logistical considerations (Richmond, Canavan et al. 2011). Costs can be considered in two ways. The first is financial cost, whereby the costs of an assessment tool in terms of salaries, disposables and materials are calculated, typically per examinee (Ram, Grol et al. 1999, Brownell, Lockyer et al. 2007, Zani, Donato et al. 2015). The second cost that is often considered is in terms of time –
faculty time, administration time, and assessment time are often measured and reported (Durning, Artino et al. 2012, Boggan, Cheely et al. 2014, Donato, Park et al. 2015).

Logistical considerations, such as the perceived practicality of an assessment tool, are often harder to measure in terms of feasibility. One method of determining this type of feasibility include faculty and trainees’ responses to questionnaires (Wagner and Lypson 2009, Garra, Wackett et al. 2011, Miskovic, Wyles et al. 2011), as well as the use of focused interviews (Tolsgaard, Arendrup et al. 2013). Other logistical measures include documenting the use of assessment tools i.e. how many work-place based assessments were completed over a time period (Alves de Lima, Barrero et al. 2007, Hanna, Mavroveli et al. 2012).

Assessment tools are often described as being feasible, yet once again feasibility is clearly a subjective concept. There are no clear guidelines as to what is feasible in terms of costs and time of an assessment tool – these issues depend on context. Clearly, increased costs are both expected and necessary in a high stakes certification examination, but should be minimized where possible in the setting of regular, frequent, in-training assessment in CBME.

2.3.4 ACCEPTABILITY

The Collins English Dictionary defines acceptable as being satisfactory, adequate or tolerable. In the education literature, the acceptability of an assessment tool is often gauged by the perceived satisfaction with its organization, effectiveness, appropriateness, duration and level of difficulty, as well as its perceived usefulness (Yeazel and Center 2004, Kneebone, Nestel et al. 2006, Wagner and Lypson 2009, Hanna, Mavroveli et al. 2012). Options for determining acceptability of an assessment tool or educational intervention include quantitative analysis with questionnaires (Hanna, Mavroveli et al. 2012), qualitative analysis of both students and faculty (Chisholm, Hart et al. 2015), as well as group discussion (Rees and Shepherd 2005, McKay, Shepherd et al. 2008). There is also some overlap between feasibility and acceptability, as both financial cost and cost in terms of time can impact the perceived acceptability of an assessment tool.
Acceptability in assessment can also relate to the number of trainees who might pass or fail an exam – this concept is discussed in further detail in 2.4.5.

2.3.5 EDUCATIONAL IMPACT

Educational impact, as described by van der Vleuten, is similar to the concept of consequences, an important source of evidence in modern theories of validity (van der Vleuten and Schuwirth 2005). Van der Vleuten described two reasons for considering educational impact as an important characteristic for determining the utility of an assessment tool - the first reason relates to the important notion that assessment impacts learning. The second, seemingly in conflict with the first, was that a focus on educational impact was required due to the scarcity of evidence to demonstrate the relationship between assessment and learning (van der Vleuten and Schuwirth 2005).

While high stakes examinations have obvious impact, such as certifying specialists as fit for practice, all assessment tools must have some education impact in order to be used. Perceived impacts of assessments, broadly speaking, include effects on training, teaching and learning (Ringsted, Henriksen et al. 2004) – examples include helping to identify learning goals, identifying deficiencies in training programs, facilitating appropriate feedback from faculty to trainees or identifying residents requiring remediation (Olupeliyawa, Balasooriya et al. 2014, Donato, Park et al. 2015). An assessment might also have negative or positive influence on learning.

2.3.6 CONCLUSION

This section has identified that assessment methods should be evaluated using contemporary concepts of validity, whereby evidence is used to support the proposed interpretation of the scores. The evaluation of an assessment program should ideally focus on the five sources of validity evidence listed above. However, while multiple sources of evidence are required, the amount of evidence necessary will vary according to the intended purpose of the instrument. While more evidence of validity is required in high stakes exams, it can be argued that less is required in the setting of regular in-training assessment in a CBME program. In this manner, instruments that are subjective or not perfectly standardized can still be used (Bhatti and Cummings
However, the assessment of competency will always require the exercise of judgment by another practitioner; for this reason a truly objective assessment is not always possible.

Van der Vleuten’s conceptual model for determining the utility of an assessment tool was published in 1996, and outlined five characteristics (van der Vleuten 1996) – the weight assigned to each criteria depends on the specific use in each specific situation. This model was not designed to be used as an algorithm, and may not cover all possible criteria (fairness, transparency). It is also important to recognize that all assessment methods require compromise depending on their purported use, especially with regards implementation and resources (van der Vleuten and Schuwirth 2005). In this way, the model can serve as an aid not just in the process of devising individual assessment methods, but also in planning overall assessment programs for a whole course, by keeping in mind what is acceptable, and the resources available.

2.4 THE ASSESSMENT OF COMPETENCE

2.4.1 INTRODUCTION

Since the inception of CBME, medical educators have been searching for comprehensive assessment tools that produce reliable and valid results (Holmboe, Ward et al. 2011). Any assessment program must be both robust and multifaceted (Holmboe, Sherbino et al. 2010). Despite widespread agreement on the importance of CBME, and more than two decades of research, a dependable set of assessment tools has not been generated (Carraccio, Wolfsthal et al. 2002, Lurie 2012).

The advantages of regular assessment are many. Frequent formative assessment ensures high quality feedback that can guide the acquisition of necessary competencies, while providing an opportunity to identify trainees with deficiencies and allow remediation (Holmboe, Sherbino et al. 2010). Regular assessment also allows program level decisions regarding trainee progression, crucial to moving away from traditional training where time was used as a proxy for competence (Carraccio, Wolfsthal et al. 2002, Holmboe, Sherbino et al. 2010). While it will be critical that research identifies acceptable and defendable levels of assessment activity, it will be
equally important to avoiding excessive assessment activity and administrative elephants (Norman, Norcini et al. 2014).

In the surgical fields, the establishment of a minimal level of competence is required in many areas, including medical knowledge, clinical decision-making, professional judgment, the performance of technical procedures, as well as the intrinsic CanMEDS roles (the six roles other than Medical Expert) (Bhatti and Cummings 2007, Frank and Danoff 2007). Real-life surgical scenarios require knowledge of facts, as well as the application of this knowledge in complex situations, requiring judgments that are often based on ill-defined parameters – there is no single test that reliably and objectively assesses all these abilities (Bhatti and Cummings 2007, Lurie, Mooney et al. 2009).

In many orthopedic training programs, exit or certification exams are used to assess competence in the role of Medical Expert, and to a much lesser degree the intrinsic CanMEDS roles. Clearly, the attainment of minimal personal, clinical, and technical skills following a residency training program is more than achieving a passing score on a single, standardized test (Bhatti and Cummings 2007). Furthermore, to this author’s knowledge, the majority of orthopedic exit examinations do not include assessments of the ability to perform technical procedures. Competency-based medical education offers a solution to this problem, by providing regular and frequent information regarding competence from multiple sources of assessment (ten Cate and Scheele 2007).

Individual competencies are difficult to evaluate separately from each other – most assessments probably measure only a single construct (Lurie, Mooney et al. 2009, Ginsburg, McIlroy et al. 2010). For this reason, it seems most likely that CMBE in the setting of postgraduate orthopedic training will be best served by a combination of assessment tools, tailored to the specific subspecialty or module, with the choice of assessment method determined by the educational context (Van der Vleuten, Norman et al. 1991). Thus, establishing competence will involve evaluations of the management of actual patients, hypothetical patients, and / or individuals representing patients, as well as the robust assessment of technical procedures (Long 2000).
At this time, the majority of rotations in the division of orthopedics at the University of Toronto use oral examinations (typically administered by the faculty supervisor), short answer written examinations, and McQs – assessment tools that suffer from the limitations that will be discussed in this literature review. The ‘Competency by Design’ white papers released by the RCPSC recommend the incorporation of key assessment tools such as encounter cards, mini-CEX, multisource feedback, logbooks, practice-based assessments, OSCE’s and other simulation methods (RCPSC 2014) - identifying the correct mix of assessment methods will be key to the future of CBME.

In orthopedics, the physician needs to be competent in clinical skills such as history-taking and physical examination, interpretation of relevant imaging, differential diagnoses, as well as the appropriate management of these conditions (Long 2000). The orthopedic surgeon must also demonstrate an ability to apply the intrinsic CanMEDS roles in clinical situations. Finally, surgeons must clearly be competent to perform technical procedures - while it may be that competency is easiest to determine in the procedural fields, where outcomes are well defined, at this time there has been limited attention to this critical area. The next sections will discuss the tools available for assessment of each of these three critical areas.

2.4.2 MEDICAL EXPERT - CLINICAL SKILLS

Options for the assessment of Medical Expert include MCQ, short answer questions, structured oral examinations (SOE), the mini-CEX, and Objective Structured Clinical Examinations (OSCEs). In certifying examinations, written examinations are used extensively, and usually take the form of pencil-and-paper examinations (Sidhu, Grober et al. 2004). The drawbacks of this method include the considerable faculty time required in order to mark these exams, and the time consuming task of data entry in order to conduct meaningful psychometric analysis. In contrast, MCQ examinations can be completed by computer software – scoring and analysis is objective. The main criticism of MCQ examinations are that this format is artificial, and that MCQ exams prompts or aids test takers by providing the correct answer (Sidhu, Grober et al. 2004).
Oral examinations involve one or more examiners asking candidates clinically based questions regarding problem-solving and operative decision-making – the oral can be structured, by utilizing examiner training, standardized scenarios and marking schemes to increase reliability (Anastakis, Cohen et al. 1991, Schubert, Tetzlaff et al. 1999, Sidhu, Grober et al. 2004). A SOE involves one or more examiners asking candidates clinically based questions regarding problem-solving and operative decision-making - previous studies in surgical residents have demonstrated that the overall reliability and inter-rater reliability of SOEs is quite high, with SOEs having good correlations with other knowledge tests such as MCQ and OSCEs (Anastakis, Cohen et al. 1991). It is also thought that direct discussion with the examiner allows the specific assessment of cognitive knowledge, problem solving abilities and communication skills.

An OSCE, consisting of a series of stations through which examiners rotate (Sidhu, Grober et al. 2004), is a popular format for examining the role of Medical Expert. OSCE’s are an important aspect of orthopedic assessment in certifying examinations both in Canada and worldwide, and OSCE scores have been demonstrated to be both valid and reliable in undergraduate, dentistry, and in postgraduate physician training (Cohen, Reznick et al. 1990, Sloan, Donnelly et al. 1996, Hodges, Regehr et al. 1997, Hodges, Regehr et al. 1998, Schwartz, Witzke et al. 1998, Brown, Manogue et al. 1999, Jefferies, Simmons et al. 2007, O'Sullivan, Chao et al. 2008). The theory and application of OSCE's in the setting of CBME is discussed extensively below in 2.4.5.

Clinical skills should also be assessed in the workplace, as ultimately this is the setting for the application of these skills as they pertain to patient care. Tools for workplace-based assessment (WBA) include the mini-CEX (Norcini, Blank et al. 2003), case-based discussion (Crossley, Johnson et al. 2011), acute care assessment tools (Johnson 2009), Longitudinal Evaluation of Performance (LEP) (Prescott, Norcini et al. 2002), and Case-based Discussion (CBD). The LEP and mini-CEX use direct observation of residents in clinical practice, typically judged by faculty over broad categories (i.e. examination and consultation skills, clinical judgment and diagnosis, technical ability and manual dexterity, communication skills, professionalism, knowledge and organization) (Norcini, Blank et al. 1995, Prescott, Norcini et al. 2002).
Judgments are made on a 9-point scale, with competence represented by being fit for safe, independent practice (Norcini, Blank et al. 1997). A meta-analysis of 13 studies performed in 2013 identified that, after use in the residency setting, the mini-CEX was able to measure differences in clinical skills performance, and had an ability to predict performance on other knowledge tests (Al Ansari, Ali et al. 2013).

### 2.4.3 INTRINSIC CANMEDS ROLES

The seven CanMEDS competencies (Medical Expert, and the six intrinsic roles of Communicator, Collaborator, Manager, Health Advocate, Scholar and Professional) have been clearly outlined in the CanMEDS 2005 Physician Competency Framework, formulated by the RCPSC (Frank 2005). Despite the widespread popularity of the CanMEDS roles and other competency frameworks, as well as a mandate to both teach and assess these competencies, the best methods of doing so remain unknown (Jefferies, Simmons et al. 2011, Zuckerman, Holder et al. 2012).

Assessment options for the intrinsic roles include ITERs (in-training evaluation reports), 360-degree assessments, and OSCEs (Zuckerman, Holder et al. 2012). A survey of a wide variety of medical and surgical program directors in Canada identified that the ITER is the most commonly used method to evaluate the CanMEDS roles (Catton 1997, Jefferies, Simmons et al. 2007, Chou, Cole et al. 2008). Comprehensive ITERs use multiple items, with descriptive anchors (Wanzel, Ward et al. 2002) – despite this, ITERs have been shown to demonstrate poor inter-rater reliability (Sidhu, Grober et al. 2004) and poor correlation with performance-based examinations such as OSCEs (MacRae, Cohen et al. 1997). The ITER is relatively subjective in nature, and some authors believe that it would benefit from the addition of more objective assessment (Jefferies, Simmons et al. 2007). However, this is not to discount the ITER, which provides useful, subjective feedback especially in regards to the intrinsic CanMEDS roles such as professionalism, communication and collaboration (Ginsburg, McIlroy et al. 2010).

The 360 degree review typically includes evaluations by multiple persons involved in a resident’s training, such as nurses and other health professionals (Sidhu, Grober et al. 2004). From a practical view, considerable resources including time are required to...
use 360 degree reviews, including care to avoid systematic bias (Sidhu, Grober et al. 2004). A systematic review conducted by Lurie et al. identified six studies examining the use of 360-degree evaluations – there was little evidence that they could be used to distinguish individual’s levels of attainment of the six ACGME general competencies, as well as limited information regarding their reliability (Lurie, Mooney et al. 2009). In one study of otolaryngology residents, rater agreement was seen between physician and non-physicians in domains such as communication and patient relationships, but not in the domain of professionalism (Roark, Schaefer et al. 2006). Another study has demonstrated that between 5 – 10 surveys were required to achieve acceptable reliability rates (Sidhu, Grober et al. 2004).

At this time, OSCEs have also been used to assess residents’ knowledge and application of the intrinsic CanMEDS roles, including communication skills (Hodges, Turnbull et al. 1996, Srinivasan 1999, Donnelly, Sloan et al. 2000, Keely, Myers et al. 2002, Hodges and Mcllroy 2003, Yudkowsky, Alseidi et al. 2004), and the demonstration of teaching skills (Scholar) (Schol 2001, Fliegel, Frohna et al. 2002). The OSCE has also been adapted to assess competencies within the Professional role (Hilliard 2000, Altshuler and Kachur 2001, Singer, Pellegrino et al. 2001), with varying degrees of success. In 2001, Singer et al. used an OSCE to assess clinical ethics, but found a low reliability with only four stations; it was felt that increasing the number of stations would be required to obtain acceptable reliability (Singer, Pellegrino et al. 2001). There is also some literature demonstrating that improved interpersonal skills are linked to higher overall OSCE performance (Sloan, Donnelly et al. 1994, Colliver, Swartz et al. 1999, Warf, Donnelly et al. 1999).

An OSCE has also been used to assess multiple CanMEDS competencies in specialties such as radiology and neonatology (Jefferies, Simmons et al. 2007, Probyn 2010). Jefferies et al. recently demonstrated that an OSCE was a valid and reliable method of simultaneously assessing multiple competencies in neonatal-perinatal medicine (Jefferies, Simmons et al. 2007). In this study, subspecialty trainees were assessed using a combination of binary checklists, 5-point CanMEDS ratings, as well as standardized patients’ (SPs) and standardized health professionals’ (SHPs) assessment of interpersonal and communication skills. Inter-station reliability was
acceptable to excellent for six of the seven Roles, with the exception of the Scholar role. Only the teaching component of Scholar was assessed – the authors recommended creating a single station to assess the competencies inherent to the Scholar role, including the ability to understand and evaluate research.

Jefferies et al. studied the use of a SOE to assess the seven CanMEDs roles (including Medical Expert) – interstation reliability was acceptable for the roles of Medical Expert, Scholar and Professional (between 0.6 and 0.8), but not for Communicator, Collaborator and Health Advocate (between 0.4 and 0.6) or for Manager (0.19) (Jefferies, Simmons et al. 2011). In comparison to their previous OSCE study, inter-station reliability was lower for all roles except for Scholar. However, costs were reduced significantly by not using SPs.

In the orthopedic division at the University of Toronto, ITERs remain the most commonly used method of assessing residents’ knowledge and application of the intrinsic CanMEDS roles. An OSCE has been shown in neonatology to be useful when assessing multiple competencies, more so than a structured oral examination. The difference may well be the use of standardized patients, which are likely to be extremely useful when attempting to study roles such as Communicator, Professionalism etc. For this reason, this thesis will examine the use of an OSCE to assess the application and understanding of multiple intrinsic CanMEDS roles, in an attempt to provide more objective evidence than that provided by an ITER.

2.4.4 MEDICAL EXPERT - TECHNICAL PROCEDURES

The ability to competently perform technical procedures is a defining characteristic of surgery, but is not measured systematically in residency (Sidhu, Grober et al. 2004, Williams, Verhulst et al. 2012). Surgical skill acquisition also has limited representation in both the CanMEDS roles (included in Medical Expert) (Group 1996) and ACGME core competencies (covered by patient care) (ACGME 1999), likely because of the significant challenges associated with its assessment (Sidhu, Grober et al. 2004). Importantly, the procedural specialties require not just the demonstration of competency in the technical skills, but knowledge of when procedures should be
performed – skills and knowledge are equally important, and one is useless without the other (Long 2000).

It is thought that most training in technical procedures is unsystematic and unstructured (Bell, Biester et al. 2009). At this time, the majority of orthopedic residents learn arthroscopic and other surgical skills by observing and assisting surgeons in the operating room, with a graduated introduction to performing surgery (Cannon, Eckhoff et al. 2006). However, it is generally accepted that this mode of learning surgical skills is not without risk to patients, and that the more practice a resident has before performing surgery on a patient, the more likely it is that he or she will perform the surgical procedure safely (LeBlanc, Hutchison et al. 2013). Furthermore, it is felt that the strategy of simple repetition without feedback is unlikely to give meaningful results in terms of skills retention (Bell, Biester et al. 2009, Schaverien 2010).

There is some evidence that current training is not providing sufficient surgical training to residents. In a report from the American Board of Surgery, program directors in general surgery in the US were asked to identify operations that were considered critically important for a graduating specialist to know (Bell, Biester et al. 2009). Of 121 must-know operations, the average general surgical trainee had more than ten experiences in only eighteen operations, less than five experiences in eighty-three operations, and less than one experience in thirty-one operations. This finding of limited exposure and experience is important, as there is literature to support the relationship of operative volume to clinical outcomes, with evidence that the performance of experts improves as knowledge acquired through training and experience increases (Patel, Groen et al. 1990, Patel, Glaser et al. 2000).

As all post-graduate medical training slowly shifts to a competency-based model, effective assessments of surgical and other technical skills after modules will become necessary. However, the best method for of assessing competence in the performance of technical procedures in postgraduate training is unknown, and is limited by both cost and access to resources. Options for the assessment of technical
skill include assessment of performance in the operating room (workplace-based assessment), and the use of simulation.

**Simulation**

While the ability to perform technical procedures will always be best assessed in the operating room, it can be difficult to standardize operations (Reznick, Regehr et al. 1997), manage time restraints (Farnworth, Lemay et al. 2001), and ensure optimal patient safety and clinical outcomes (LeBlanc, Hutchison et al. 2013). For these reasons, simulation is being increasingly used as an assessment tool (Cannon, Eckhoff et al. 2006, Howells, Gill et al. 2008, Alvand, Auplish et al. 2011, Butler, Olson et al. 2013, LeBlanc, Hutchison et al. 2013).

Certainly, the surgical community seems to support the use of simulation in training and assessment. A recent survey of orthopedic program directors and residents identified that a significant number believed surgical skills simulation should become a required part of training, while 76% of program directors surveyed reported having a surgical skills laboratory (Karam, Pedowitz et al. 2013). There is also now a requirement by the Residency Review Committee in Surgery that all residents must have training outside of the operating room (ACGME). This requirement has been supported by the development of a national skills curriculum in simulation, through a joint venture of the Association for Program Directors in Surgery and the American College of Surgeons (Scott and Dunnington 2008, Korndorffer, Kasten et al. 2010).

The advantages of using simulators for the training and assessment of technical procedures are many. Simulation allows the opportunity for residents to attempt complex surgical tasks, without risk to patients. Surgical simulation also allows the acquisition of technical skills within a clinical context, and tailored to skill level in a safe environment (Schaverien 2010). When performing technical procedures in the laboratory, residents have the opportunity to make independent decisions, as well as take responsibility for and managing the consequences of their surgical actions, an opportunity rarely available in the operating room. Finally, simulation gives staff surgeons’ insight into each resident’s knowledge and skills deficits, prior to performance of procedures in the operating room.
A systematic review and meta-analysis in 2011 examined the outcomes of technology-enhanced simulation training for health professional learners – 609 studies were identified, demonstrating that simulation training was associated with significant effects on knowledge and skill outcomes, and moderate effects on patient related outcomes (Cook, Hatala et al. 2011). However, the move to using simulators as methods of assessment is, according to some authors, both problematic and controversial (McDougall 2007). Arguments against the use of simulation in assessment are that simulators were developed to be used as training aids, not as assessment tools – validity evidence for the use of scores in this setting must be obtained (Korndorffer, Kasten et al. 2010). Furthermore, simulator performance has typically been compared with year of training or operative experience rather than operative performance, which is the ideal criterion (Korndorffer, Kasten et al. 2010).

The use of simulation to assess the performance of technical procedures is discussed further in 2.4.6.

**Workplace-based Assessment**

Competency in a controlled environment does not ensure competent workplace performance (Rethans, Sturmans et al. 1991, Crossley, Johnson et al. 2011). Although simulation will certainly play an increasingly important role in competency assessment over time, the direct observation of learners providing care will remain a cornerstone of the assessment process (Iobst, Sherbino et al. 2010). Certainly, when applying Miller’s pyramid, it has been argued that while simulation can be used to demonstrate that a trainee ‘knows’, and ‘knows how’, a workplace-based assessment (WBA) allows a resident to ‘show how’ and ‘do’ (Ali 2013). This is debatable however, as clearly surgical simulation demands that a resident shows how.

It is important to distinguish between performance-based assessment (PBA) and workplace based assessment (WBA). PBA is a means of measuring a student’s ability to apply content knowledge to critical thinking, problem solving, and analytical tasks – nearly all assessments fall into this category. WBA involves assessment of performance (integration of clinical knowledge and skills into judgments and decision making) in everyday clinical practice.
It is believed by many authors that the primary purpose of WBAs is as a formative tool, used to guide learning (Beard, Rowley et al. 2009, Mitchell, Bhat et al. 2011, Saedon, Salleh et al. 2012, Ali 2013, Shalhoub, Vesey et al. 2014). However, surveys of both trainees and surgeons indicate dissatisfaction with the mandatory institution of these assessments tools, owing to issues ranging from time constraints, low inter- and intrarater reliabilities, limited evidence of validity, insufficient assessor training, inappropriate use of WBAs by trainees, and concern that the use of WBAs does not improve training (Albanese 2000, Williams, Klamen et al. 2003, Miller and Archer 2010, Bindal, Wall et al. 2011, McGill, van der Vleuten et al. 2011, Eardley, Bussey et al. 2013, Govaerts, Van de Wiel et al. 2013, Pereira and Dean 2013, Shalhoub, Vesey et al. 2014, Torsney, Cocker et al. 2015). It is also not always possible to sample all the critical elements of a curriculum using WBA. Despite this, the use of WBAs as a component of competency assessment continues to rise.

In 2013, Eardley et al. reported on the initial experience of WBAs in the UK over a 12 month period (17,904 case-based discussions, 11,132 PBA, 2,249 direct observation of procedural skills (DOPs), and 14,950 mini-CEXs) (Eardley, Bussey et al. 2013). Emerging themes were that trainees were using the assessments inappropriately, often waiting until they would pass before undertaking the assessment, as well as being selective when choosing the faculty to perform the assessments. In the study by Eardley, while the global WBA scores correlated with the stage of training, summary scores for many of the individual assessments did not. There was also a clear paucity of low scores. The major recommendation stemming from this study was for increased faculty training in the use of WBAs.

In 2011, Mitchell et al. reported on a retrospective analysis of WBAs (mini-CEX, case based discussions, DOPS, and mini-peer assessment tool assessments) in 1,646 UK trainees between 2005 and 2009 – while there was an association between those reported as having training difficulties and lower scores on the mini-CEX and case based discussions, overall there was a weak predictive value of lower WBAs scores (Mitchell, Bhat et al. 2011). The authors concluded that while WBAs have value as formative assessments, they should not, in isolation, be used to certify competence. Mitchell et al. hypothesized that part of the problem was that WBAs are trainee led,
with trainees in difficulty limiting exposure to further evidence of incompetence – the authors also acknowledged the subjective methods by which residents in difficulty were identified (Mitchell, Bhat et al. 2011).

Psychometric evaluations of WBAs have demonstrated significant effects related to raters, with tremendous variability seen in the rating of trainee performances (Kogan, Holmboe et al. 2009, Crossley, Johnson et al. 2011, Jelovsek 2015). It is thought that these differences relate to the way assessors value different aspects of performance (Kogan, Conforti et al. 2015). Using both qualitative and quantitative analysis to analyze the nature of rater judgments in WBAs. Govaerts et al. identified substantial levels of rater idiosyncrasy, independent of rater experience - it was recommended that the most credible method of overcoming these issues was the use of multiple assessment tools used by multiple health care providers (Govaerts, Van de Wiel et al. 2013).

Despite these limitations, the Royal College of Surgeons and the Surgical Specialty Association in the UK have adopted procedure-based assessment (PBA) (using a combination of checklists and global ratings) as the principal method of assessing surgical skills (Beard, Rowley et al. 2009). While these PBAs are used to guide the learning of index procedures and to provide constructive feedback, these forms (along with logbooks), are being used by Program Directors to make summary judgments about the competence of trainees to perform procedures (Beard, Rowley et al. 2009). At this time, there is minimal literature to support the use of PBAs in this manner. Beard et al. demonstrated that a PBA was reliable when used to assess performance of a sapheno-femoral disconnection procedure, and that performance on the PBA correlated with both performance on a simulator and previous surgical exposure (Beard, Jolly et al. 2005).

In an attempt to reduce the subjectivity of WBA, the UK Surgical curriculum program has introduced the concept of independence. An entrustability scale for assessing technical and non-technical abilities was developed, whereby the assessor must judge whether the trainee is; (i) unable to perform the procedure, or part observed, under supervision, (ii) able to perform the procedure, or part observed, under supervision, (iii)
able to perform the procedure with minimal supervision (needed occasional help), or
(iv) competent to perform the procedure unsupervised (could deal with complications
that arose) (Beard, Marriott et al. 2011). In a study by Beard et al. using this scale,
different judges were able distinguish between trainees with good reliability (0.76)
(Beard, Marriott et al. 2011). Other studies have also demonstrated that assessment
scales aligned to increasing sophistication and independence can provide a more valid
construct for clinician assessors than conventional scales (Crossley, Johnson et al.
2011).

Despite the issues highlighted above, it is clear that WBA must remain an important
component of CMBE, especially as WBA is an acceptable way to reassure the public
that medical trainees are competent at patient care (Jelovse 2015). Fortunately, there
are proposed methods of improving the performance of WBAs. While studies of
individual WBA tools have shown poor reliability (McGill, van der Vleuten et al. 2011),
unless eight to twelve assessments are performed (Wilkinson, Crossley et al. 2008),
WBA toolboxes have shown more acceptable reliability. Using G-studies to evaluate a
WBA toolbox in the Netherlands (953 residents with 12,779 WBAs), Moonen-van Loon
et al. identified that a reliability coefficient of 0.8 was obtained using eight mini-CEXs,
nine DOPS, and nine MSFs (Moonen-van Loon, Overeem et al. 2013). The authors of
this study concluded that it was possible to make summative decisions, if
combinations of WBAs were used.

2.4.5 OSCE

Theory

First developed in the late 1970’s, an OSCE consists of multiple, timed stations at
which each candidate is faced with a discrete clinical task. Examinees are evaluated
in an objective and structured way (either by an experienced clinician, a standardized
patient, or both); both the task and the rating of performance are standardized (Harden
2008).
Especially suited to testing aspects of the role of Medical Expert (Cohen, Reznick et al. 1990, Hodges, Regehr et al. 1998, Schwartz, Witzke et al. 1998), an OSCE is extremely useful as a performance-based assessment, providing a means of assessing a broad array of examinees, from medical students to residents and experienced doctors (Brannick, Erol-Korkmaz et al. 2011). Widely used to assess clinical competence (Gaur and Skochelak 2004), OSCE assessments can ensure examinees can generate a diagnosis based upon history and examination, interpret imaging, and propose management plans. In this manner, OSCEs are used to test a broad range of clinical, technical, and practical skills (Newble 2004).

Creating an OSCE that produces reliable and valid scores involves a series of accepted steps - this process generally begins with a focus on content, involving establishment of a defined curriculum, a clear set of learning objectives, and validation by an expert panel (Newble 2004). After defining the range of conditions that the candidate needs to be competent at managing, a blueprint is created with a range of tasks for each condition and used to ensure wide sampling for each assessment (Newble 2004). Objectivity is achieved through this task standardization, but is also related to the scoring systems used (Harden and Gleeson 1979).

Checklists were originally introduced in combination with global rating scales (GRS) in order to enhance inter-rater reliability (Newble 2004), a combination that has been demonstrated to produce valid results (Hodges and McIlroy 2003) (Reznick, Regehr et al. 1998, Hodges, Regehr et al. 1999). However, it is thought that global ratings have a higher degree of reliability, as well as possibly being more sensitive to levels of training. Regehr et al. showed that global ratings completed without the use of a task-specific checklists were as valid and reliable as global ratings used in association with a checklist, suggesting that completion of a checklist does not influence the use of a GRS (Regehr, MacRae et al. 1998). It is also thought that relying on checklists alone results in possible over-objectification of performance, whereas less quantifiable aspects of performance can be taken into account using a GRS (Norman, Van der Vleuten et al. 1991, Van der Vleuten, Norman et al. 1991). Furthermore, using a global rating scale allows highly organized and succinct individuals to be rewarded,
compared with candidates who score marks on a checklist, but are less ordered and overall perform less strongly (Hodges, Regehr et al. 1999).

In 2011 Brannick et al. conducted a systematic review of the reliability of OSCE scores across 39 studies – the overall alpha across stations was 0.66 (0.62 – 0.70), while the overall alpha within stations and across items was 0.78 (0.73 – 0.82) (Brannick, Erol-Korkmaz et al. 2011). In this review, increased reliability was associated with increased numbers of stations, increased number of examiners per stations (one versus two), while decreased reliability was associated with the assessment of interpersonal skills. No significant difference was seen in the reliability depending on whether Likert scales or checklists were used, or whether faculty or SP performed the rating. The mean of the generalizability coefficients was lower than the mean of the alpha coefficient (mean of 0.49 v 0.62).

While increasing the number of stations is generally considered to increase the reliability of an OSCE, there are limits, as stations can be expensive. Furthermore, it is not a given that increasing stations will increase reliability, as other factors affect reliability – the reliability of an OSCE is also dependent on careful sampling over multiple topics, and influenced by examiners (van der Vleuten and Schuwirth 2005). Increasing the number of raters can also increase reliability, although it is important to decide if increasing the number of stations is a more appropriate use of a valuable resource such as faculty time (Govaerts, van der Vleuten et al. 2002). Importantly, it is known that the most crucial aspect of increasing intra- and inter-examiner reliability is experience and training of the examiners.

**OSCE in Orthopedics**

Until recently, there was little in the way of published literature on the validity and reliability of OSCE scores in orthopedics, either as a high stakes exit examination, or as a component of in-training assessment. One recently published paper used an OSCE to assess the history-taking and physical examination skills of orthopedic residents, finding that the OSCE produced reliable scores that served to highlight deficiencies in resident knowledge (Beran, Awan et al. 2012). Chen et al. videotaped an interaction between a standardized patient with a femoral neck fracture, and
orthopedic staff, fellows, and senior residents, studying the intra- and inter-rater reliability (Chen, Lee et al. 2013). When using a dichotomous pass/fail rating, the inter-rater reliability was >0.8, but only a fair reliability was seen when using checklists alone. Vivkananda-Schmidt et al. videotaped medical students performing joint examination skills – senior residents acted as raters while a consultant rheumatologist reviewing the videotapes (Vivekananda-Schmidt, Lewis et al. 2007). There was between an 84 and 87% agreement of the classification of examination grades. Finally, Mohtadi et al. studied inter-rater reliability in a certifying examination for sports medicine diplomas; using two observers at 4/20 stations, the reliability co-efficient ranged from 0.85 to 0.99 (Mohtadi, Harasym et al. 1995).

Recently, the author of this thesis examined the reliability and validity of an OSCE as an in-training assessment tool for orthopedic residents after a sports medicine module (Dwyer, Theodoropoulos et al. 2013) (Appendix C). Six stations were written, with each station based on common clinical scenarios from a predetermined curriculum – sampling of scenarios was chosen to mimic the clinical load in a typical orthopedic sports medicine practice (i.e. two shoulder cases, two knee cases, two other (hip, ankle, elbow). The primary purpose of the OSCE was to test clinical skills in the CanMEDS role of Medical Expert, focusing on six domains; history-taking, examination, image interpretation, clinical decision-making, consent, and knowledge of surgical procedures. Each station (10 minutes duration) was based upon a single clinical scenario - no standardized patients were used, but the OSCE nomenclature was maintained.

A binary station-specific checklist was combined with a single 5-point global rating (1 = clear fail, 2 = borderline, 3 = pass, 4 = good pass, 5 = excellent pass) (Reznick, Regehr et al. 1998, Hodges, Regehr et al. 1999, Hodges and McIlroy 2003) (Hodges, Regehr et al. 1997, Pell G 2006, Pell, Fuller et al. 2010). Examiners were instructed that a pass mark should be equivalent to a pass at an orthopedic certifying examination level, and not related to year of training.

A total of 43 residents (with a mix of sports medicine training and experience) from the University of Toronto volunteered to sit the OSCE (two-thirds of all residents) over a
period of two weeks. Data analysis showed that the Cronbach’s alpha was 0.91 for the 6 stations, well within acceptable limits for a medium stakes OSCE. In no station did Cronbach’s Alpha if Item deleted rise, indicating individual station reliability, while the Corrected Item Total Correlation Coefficients for each of the six stations were all very acceptable. Total test scores (average of all station-level checklist scores) were significantly higher for PGY-5 residents than PGY-1, 2 and 3. One-way analysis of variance (ANOVA) of the scores for each of the domains (apart from history) showed a statistically significant effect of postgraduate year of training, while post hoc analysis showed that the domains of examination, image interpretation, clinical decision-making, and surgical techniques had the most striking correlation with year of training. Finally, a correlation was seen between the results of the sports medicine OSCE and each resident’s Orthopedic In-Training Examination (OITE) total test scores and the sports medicine subscores.

The results of this study demonstrated that a sports medicine OSCE produced scores that were sufficiently reliable to be used as an in-training assessment of competence. Furthermore, the OSCE scores had evidence of validity, as the OSCE was able to distinguish between different levels of training, within the majority of knowledge domains, as well as demonstrating a correlation between increasing sports medicine experience and an increase in OSCE scores. Naturally, increasing experience is closely related to advanced year of training, although, surprisingly, having undertaken a recent sports medicine rotation (run along a competency-based model) did not have a significant effect on performance, although there was a trend to improved scores. However, the ability to comment on this effect was limited in the study due to the low number of residents (11) who had been exposed to the new curriculum.

Importantly, the results of this study demonstrated that it was possible to create an OSCE that produced reliable scores using only six stations. While it is commonly accepted that increasing the number of stations increases the overall reliability, it was concluded that focusing on a narrow range of content (orthopedic sports medicine) enhanced the reliability. For this reason, it was decided to continue with the OSCE format to assess the clinical competence of residents after a competency-based orthopedic sports medicine rotation.
It is recognized that the definition of an OSCE is variable (Aggarwal and Darzi 2006), (Harden and Gleeson 1979, Cohen, Reznick et al. 1990, Newble 2004). A typical OSCE consists of a series of stations through which examinees rotate – at each station an examinee is asked to perform a clinical task (Sidhu, Grober et al. 2004). Stations may or may not involve a standardized patient (Hodges, Regehr et al. 1999). Certainly, the lack of SP’s can be criticized (Newble 2004). One of the major advantages of the OSCE format described above that did not use SP’s was both its relative simplicity to run, and its minimal costs. The main focus of the sports OSCE has been on the management of clinical problems – while SPs are useful for history taking, the author of this thesis believes that this skill can be adequately assessed without an SP. It is also believed by this author that assessing examination skills on SPs who do not have pathological findings is of limited value to residents, especially as they become more advanced.

An alternative option would be to use real patients rather than SPs – however, SPs have become the standard in OSCEs for many reasons including their ready availability, their ability to reliably present the same information over and over, as well as avoiding potential mistreatment of patients (i.e. exposure to inappropriate comments, poor examination techniques) (Barrows 1993). After reviewing the literature and experimenting with different OSCE models, we determined that, based on cost and feasibility, our OSCEs would employ role-playing examiners in some cases but not standardized or real patients - in stations that involve history-taking, residents will take a history from the examiner.

Standard Setting in OSCEs

The goal of the in-training OSCE in the setting of CBME is to establish a minimal level of competence, the single most important aspect of which is the determination of the actual standard to be applied. While such a standard is always arbitrary, it must be defendable (Cusimano 1996). In an OSCE, a cut-off score is required to set a pass/fail standard by using standard setting methods (Cusimano 1996).

Standard setting may use either a relative (norm-referenced) or absolute (criterion-referenced) approach (Newble 2004). **Norm referencing** determines that a fixed
number or percentage of candidates will pass, and tends to be used when a limited number of places are available i.e. entry to a course (Norcini 2003). An obvious disadvantage of this system is that some candidates will pass and others will fail regardless of how well they perform – even is all candidates perform very well on an exam, only a few can be selected (Muijtjens, Hoogenboom et al. 1998, Pell 2006, Schoonheim-Klein, Muijtjens et al. 2009).

Standards are criterion-referenced when the pass/fail point is set on the basis of what the student should know; this is the preferred option when establishing competence in the setting of CBME (Turnbull 1989, Norcini 1999, Schoonheim-Klein, Muijtjens et al. 2009, lobst, Sherbino et al. 2010). Criterion referencing implies that competence should depend on an acceptable level of performance, and not be adjusted to regulate the proportion of persons passing the test. More than 30 different criterion-referenced standard setting methods have been described (Cusimano 1996).

**Test-centred methods**

Criterion standard-setting methods can be divided into judgmental / test-centred (based on item or station characteristics) methods, and empirical / examinee-centred (requiring test data to aid in standard setting) methods, or combinations of the two (Hambleton 1980, Jaeger 1989). In a test-centred method, the test content is reviewed by judges and a standard set; in an examinee-centred method the pass/fail standard is determined by judging the performance of individual students relative to a standard based on overall performance (Angoff 1971, Schoonheim-Klein, Muijtjens et al. 2009).

An example of a test-centred method is the Angoff method (Angoff 1971). In this method, judges are asked to review each item in a test and to estimate ‘the probability that a ‘minimally acceptable’ or borderline candidate would answer an item correctly’ (Angoff 1971). The judges’ assessments of an item are then averaged to determine the probability of a correct response for an item. The advantages of this system include its relative simplicity, the fact that the pass/fail standard can be set before the OSCE is undertaken (Schoonheim-Klein, Muijtjens et al. 2009), and that this method has been shown to have sufficient intra-panel and inter-panel reproducibility (Norcini
The disadvantage is that often judges tend to set the Angoff standard too high – judges find it difficult to focus on borderline candidates and often consider average students, thus inflating the pass mark (Pell 2006). The reality check was introduced to address this issue, whereby the Angoff is performed both before and after the examination (Kramer, Muijtjens et al. 2003) – in this manner judges can revise their judgments in the light of the OSCE scores of real students, and with group discussion (Morrison, McNally et al. 1996).

The modified Angoff method was also designed to address this issue. In the modification judges ask the question ‘would a borderline candidate be able to answer this item correctly?’ Items are then assigned as yes=1 and no=0, and the pass point is calculated by averaging the scores. This method is thought to be much easier than estimating the proportion correct (Downing 2006). The various modifications of the Angoff method have also been studied, including a three-level Angoff variant (in which judges decide the probability that a borderline candidate would answer each item correctly (yes/no/maybe), in comparison with the modified Angoff method - the scores from the latter were found to be more credible and reliable (Jalili, Hejri et al. 2011).

Both the number of judges required, and the expertise of judges required for the Angoff method is debated in the literature. Verhoeven et al. studied the application of the Angoff procedure in the setting of an undergraduate exam using new medical graduates as judges, concluding that this was an appropriate standard setting method (Verhoeven, van der Steeg et al. 1999). In a follow-up study, Verhoeven et al. compared the standard setting of item writers to the new medical graduates from their previous study; they concluded that the item writer’s passing score was less credible, and that high numbers (39) of judges were required (Verhoeven, Verwijnen et al. 2002).

Verheggen et al. demonstrated considerable variation between judges, especially when judges had less expertise in certain item areas (Verheggen, Muijtjens et al. 2008). The authors recommended careful judge selection, and that judges should be capable of answering all items correctly, as well as raising the issue of whether answer keys should be provided. For this reason it is commonly recommended that sufficiently
high number of judges are used, that judges are well informed about their task, and that discussion is allowed (Williams, Klamen et al. 2003). It has also been suggested that a mix of judges may be advantageous, including generalists, specialists, with a mix of gender and age (Norcini 2003).

The Ebel method involves judges rating each item in two dimensions; difficulty (easy, medium, hard) and relevance (important, acceptable, questionable) (Ebel 1979). Using a matrix, categorized by relevance and difficulty, judges assign items to a relevance and difficulty cell, and then estimate what proportion of marginal students would accomplish those items (Yudkowsky, Downing et al. 2008). The number of items in each cell is then multiplied by the estimated percentage, the product in each cell summed, then divided by the total number of items (Cusimano 1996). Variations include determining item difficulty based on actual data (Downing, Tekian et al. 2006). The advantage of this system is that it is easy to compute, and easy for judges to understand, however it can be difficult to keep the two dimensions separate (Cusimano 1996). Modifications include removing the ‘hard’ component of difficulty, as well as removing the ‘questionable’ relevance category (Cusimano 1996).

In 2008, Yudkowsky et al. compared the Angoff method with the Ebel method, demonstrating that the Angoff method generated lower cut scores than the Ebel Method, and was much preferred by the judges (Yudkowsky, Downing et al. 2008). In 2003, Downing et al compared four absolute standard setting method in the setting of a written examination, including the Direct Borderline method (really a modified Angoff method), the Ebel method, Nedelsky and the Hofstee method (Downing, Lieska et al. 2003). The modified Angoff method produced acceptable and reproducible passing scores, and was noted to be extremely feasible in terms of time commitments. While the Ebel method had a lower SEM and good reproducibility, both the Ebel method and the Hofstee method produced the lowest passing scores and highest passing rates, and were deemed to be generally unacceptable.
Examinee-centred methods

An example of an examinee-centred method used by medical schools (e.g. the University of Ottawa and the University of Otago), and by the Medical Council for Canada is the Borderline Group (BG) method (Dauphinee, Blackmore et al. 1997, Wilkinson, Newble et al. 2001, Wood, Humphrey-Murto et al. 2006). This method involves examiners in each station giving a global rating of each student’s overall performance independent of the checklist mark (Kaufman, Mann et al. 2000, Wilkinson, Newble et al. 2001). The mean of scores for all candidates identified as borderline on the rating scale becomes the pass mark for the station, with the pass mark for the whole OSCE calculated by averaging each station’s borderline score (Newble 2004). One problem that can occur is the situation where one examiner identifies an increased number of borderline candidates, thereby having a greater influence than other examiners (McIlroy 2000). This is often overcome in the situation whereby larger number of examiners observe a large number of students, increasing reliability (Pell 2006). It has also been claimed that setting a passing score with the BG method is sample dependent, and not based on a rigorous criterion of competence (McIlhenny and Orr 2002).

A modification of the BG method is the Borderline Regression method (BR), which uses a linear regression approach to set cut scores, potentially avoiding unequal weighting of examinees rated as borderline, especially in small scale OSCEs (Kramer, Muijtjens et al. 2003). This method regresses all of the examiners’ checklist scores onto their global ratings to produce a linear equation, allowing the midpoint of the borderline group to be determined (Wood, Humphrey-Murto et al. 2006). Comparing the BG and BR methods, Wood et al. found that that there were relatively small differences in the cut scores and pass rates, although the BR method had a slightly lower cut score for most stations, and was more accurate (having a smaller 95% confidence interval) (Wood, Humphrey-Murto et al. 2006). However, the authors thought that the modified BLR method might potentially be more appropriate in the setting of small sample sizes.
Comparative studies

Applying the BG method to a medical student OSCE in 2004, Kilminster found this method to be feasible and credible, as well as time efficient (Kilminster and Roberts 2004). A comparison of the pass marks obtained with the modified BG method, the BG method, and the Angoff at three medical schools showed more consistency with the BG and modified BG methods, although it was noted that passing standards at different medical schools was not equivalent (Boursicot, Roberts et al. 2007). In 1991, Woehr et al compared seven methods to establish cutoff scores in an MCQ exam, including BR methods and the Angoff method (Woehr 1991). While the various methods resulted in slightly different cut-scores, they all fell within the Standard Error of Measurement (SEM) of the exam, and differences in mean performances of persons passing with each cut-score was minimal. The use of the SEM as the unit of analysis was thought to be an important aspect of comparing accuracy of each method, as the SEM is based on the reliability of the test (Cusimano 1996).

In a variation of the BG method, McIlroy examined the effect of cut scores using the traditional BG method, versus calculating examiner specific mean checklist scores for the borderline group, then computing the average of these means across examiners, in an effort to weight all examiners opinions equally (McIlroy 2000). A high degree of inconsistency was noted; in some instances one examiner would identify 90% of candidates as borderline, while another identified no borderline candidates – this corresponded to a high discrepancy in cut scores. However, little difference was seen in cut scores and pass-fail decisions.

Both the Angoff and BG / BR methods have been applied in the setting of an OSCE (Morrison, McNally et al. 1996, Kaufman, Mann et al. 2000, Wilkinson, Newble et al. 2001, Kramer, Muijtjens et al. 2003, Kilminster and Roberts 2004, Wood, Humphrey-Murto et al. 2006, Schoonheim-Klein, Muijtjens et al. 2009). Examining standard setting in postgraduate general practice training, Kramer et al. compared the BR method to the Angoff method, finding that the BR method was more credible (high correlation of test difficulty and passing scores) and reliable (generalizability theory) than the modified Angoff (Kramer, Muijtjens et al. 2003). Schoonheim-Klein et al. also
compared the BR method to the Angoff method in a dental student’s OSCE; the BR method had the highest pass rate, and the highest reliability (lower root mean square error / RSME) – the authors believed that this method was the optimal choice for OSCEs in health education (Schoonheim-Klein, Muijtjens et al. 2009). In 2000, Kaufman et al. compared the Angoff and the BG procedures, and found that both provide reasonable and defensible results in the medical school setting (Kaufman, Mann et al. 2000).

With regards to the Angoff method, the RMSE has been used to estimate the minimum number of judges – by estimating that a 1% shift in passing score (RMSE 0.51) changes the failure rate by approximately 2.5%, 10 judges would be required to rate 200 items or more (Verhoeven, van der Steeg et al. 1999). Using average variances, Fowell et al. calculated that the minimum number of judges to obtain a RMSE of less than 2% is 10 or more judges before discussion, and 6 or more after discussion (Fowell, Fewtrell et al. 2008). This study indicated that a discussion phase improved the reliability, and reduced the number of judges required. It was also suggested that, in the setting of judges working independently, the median rather than the mean should be used to minimize hawk and dove effect, and that the effect of using a combination of expert and non-expert judges be studied. In a high stakes exam, it is thought that six to eight judges are a minimum (Brennan 1980)

**Summary**

In the context of regular assessment as part of CBME, it will not be plausible to use an examinee- centred method such as the BG to determine competence. In the current training model at the University of Toronto, only 5 – 6 residents undertake the sports medicine module every three months; because of these low numbers a test-centred method, such as the modified Angoff or Ebel method will likely be the best method of establishing a cut-score, the proxy for a designation of competence.

It is also unknown what are the best characteristics of the judges involved in these methods – who will determine the required level of competence? At this time, the acceptable level of knowledge that a resident should have after a sports rotation is
unclear; the risk is that the exam writers in this module, all of whom are experts in the field of sports medicine, may set the pass mark too high. It may be more acceptable to use general orthopaedic surgeons, or surgeons whose subspecialty interests or training lie in another area. However, these judges may set the pass mark too low. This is where the subjective nature of determining competency must be acknowledged, and requires further investigation.

Finally, the current curriculum at the University of Toronto has some rotations (including sports medicine) that have both junior and senior rotations. Similar to the arthroplasty and trauma rotations, there are insufficient resident numbers to feasibly run junior and senior OSCEs – for this reason, faculty must consider the issue of the standards they expect from junior and senior residents. Is it reasonable to require the same level of competence from each, or should there be different expectations of junior and senior residents?

As a result, I propose studying the use of the modified Angoff method to set different standards for junior and senior residents, using the BG / BLR methods as ‘gold standard’ to establish credibility. I have chosen the modified Angoff method as it seems relatively simply for judges to understand and apply, and there is a body of evidence to support its use. Three groups of judges will be used (sports surgeons, general surgeons, and newly graduated orthopaedic surgeons) in order to understand the ideal characteristics of judges applying this method. Finally, I will attempt to establish if it is conceivable, plausible, and acceptable to have differing levels of competence, depending on year of training of residents.

At this time, standard setting methods following in-training assessment in the setting of CBME have not been studied. This is clearly important - the purpose of an in-training assessment is to determine a minimal level of competence in orthopedic residents. Without an ability to set a cut-score, who passes and who fails is unknown - standard setting methods are crucial to identify the resident in difficulty, who is far below their peers in terms of performance, and would benefit from remediation.
2.4.6 OSATS

The Objective Structured Assessment of Technical Skills (OSATS) was developed at the University of Toronto (Martin, Regehr et al. 1997, Reznick, Regehr et al. 1997), as an objective measurement format used to assess the technical competency of surgical trainees. An OSATS consists of a multi-station, OSCE-like examination, in which candidates perform a series of standardized surgical tasks (Sidhu, Grober et al. 2004). Held either with high-fidelity models (cadavers, animals) or low-fidelity bench models (sawbones), assessment usually takes the form of a task-specific checklist in combination with a GRS. Often these assessments are combined with a final rating such as demonstrates proficiency or requires further practice.

At this time, it is thought that an OSATS is the most accepted ‘gold standard’ for objective skills assessment (van Hove, Tuijthof et al. 2010). However, Van Hove et al. performed a systematic review of methods of objective skills assessment, and concluded that the while OSATS scores had good reliability and evidence of validity in the laboratory setting, at this time there is limited evidence for its use in the operating room (van Hove, Tuijthof et al. 2010). The rising importance of the OSATS model is reflected by the American Board of Surgery, which now requires successful completion of the Fundamentals of Laparoscopic Surgery Program (an OSATS style exam) prior to board certification (Dulan, Rege et al. 2012).

When considering using an OSATS in the setting of simulation, it is critical to decide which type of model to use. Cadavers are thought to be the gold standard for simulation training and assessment (Reznick, Regehr et al. 1997, Anastakis, Regehr et al. 1999), but require advanced facilities. It is also accepted that there are significantly increased costs associated with cadaveric models compared to low fidelity models (Zendejas, Wang et al. 2013). The sports medicine program at the University of Toronto reserves cadavers for annual arthroscopy teaching days - these skill sessions provide an incredible hands-on learning opportunity for residents, under faculty supervision. However, because of cost, it is believed that cadavers are not a feasible option for regular assessment of technical procedures within a CBME model of orthopedic training.
Virtual reality (VR) simulation is commercially available for knee and shoulder arthroscopy – studies have demonstrated an ability to distinguish between operators of differing skill level, an ability to demonstrate improvement in skills over time, as well as improvement of knee and shoulder arthroscopy skills in the operating room (Pedowitz, Esch et al. 2002, Srivastava, Youngblood et al. 2004, McCarthy, Moody et al. 2006, Gomoll, O'Toole et al. 2007, Gomoll, Pappas et al. 2008, Rebolledo, Hammann-Scala et al. 2015). The downsides of VR are cost (including maintenance), as well as limited haptic feedback or physical sense of resistance. Currently, computer-simulated arthroscopy models test the ability of a subject to locate and probe a simulated target multiple times within a knee or shoulder joint (McCarthy, Harley et al. 1999, McCarthy, Moody et al. 2006) - at this time VR cannot be used to simulate more complex sports medicine procedures (i.e. insertion of anchors in the humeral head or glenoid, suture passage, or the drilling of ACL tunnels). LeBlanc et al. studied this issue by comparing performance of surgical fixation of an ulna on both a virtual simulator and on sawbones – while the VR model showed promise, it did not achieve the standard of the sawbones model (LeBlanc, Hutchison et al. 2013).

The advantages of low fidelity, dry models are many, including relative ease of preparation, reduced cost compared to cadavers, as well as the ability for unsupervised practice (Butler, Olson et al. 2013). The majority of comparative studies in other surgical specialties have demonstrated that, overall, low-fidelity simulators are as effective as high-fidelity simulators with regards to the acquisition of surgical skills (Matsumoto, Hamstra et al. 2002, Grober, Hamstra et al. 2004, Chandra, Savoldelli et al. 2008, McDougall, Kolla et al. 2009, Zendejas, Wang et al. 2013). One example is the study by Anastakis et al., where first year general surgery residents were trained on a series on tasks – one group used bench models, while the other used cadaveric models (Anastakis, Regehr et al. 1999). One week later each group performed the task on cadaveric models, with no significant difference in performance between the two groups.

Studies in the field of orthopedics support the use of dry models. In 2008, Howell et al. used a randomized trial to demonstrate that simulation training on a knee sawbone
model improved performance in the operating room when performing diagnostic knee arthroscopy (Howells, Gill et al. 2008). Other studies have shown that the use of low-fidelity models improves the skills of medical students learning to perform triangulation in shoulder and knee arthroscopy (Alvand, Auplish et al. 2011, Butler, Olson et al. 2013), as well as perform meniscal repair (Alvand, Logishetty et al. 2013).

A major advantage of dry models over virtual reality is the ability to perform complex procedures. Recently, the author of this thesis conducted a study using dry models to assess performance of ACL reconstruction (Appendix D) (Dwyer, Slade Shantz et al. 2015). This study demonstrated high internal consistency, acceptable inter-rater reliability, and evidence of validity (an ability to differentiate between different years of training, and a correlation with surgical experience). The costs were also shown to be acceptable in the setting of regular assessment.

Regardless of the model used for an OSATS, reliable and validated global rating scales are required in order to allow assessment. In general surgery and open procedures, the global rating of operative performance developed by Reznick et al. has been shown to be valid and reliable, and is commonly used (Reznick, Regehr et al. 1997, Regehr, MacRae et al. 1998). In 2013, Koehler et al. published the Arthroscopic Surgical Skill Evaluation Tool (ASSET), a global rating scale with eight domains, including camera and instrument dexterity, flow of the procedure and quality of the procedure (Koehler, Amsdell et al. 2013). Each domain is rated from 1 to 5, based upon the Dreyfus model of skill acquisition (Novice, Competent, Expert). Other global rating scales have also been developed for use in the assessment of arthroscopic procedures. Insel et al. developed the Basic Arthroscopic Knee Skill Scoring System (Insel, Carofino et al. 2009), while Slade Shantz et al. published the Objective Assessment of Arthroscopic Skills (Slade Shantz, Leiter et al. 2013), both of which were reliable and showed reasonable correlation with arthroscopic experience. The major advantage of ASSET global rating over these over rating scales, in the opinion of this author, are two-fold; firstly, the ASSET was designed to be generalizable for all arthroscopic procedures (not just knee arthroscopy), and secondly because it is anchored using the Dreyfus model of skill acquisition.
At this time, OSATS have not been used in the setting of regular assessment in CBME programs. Recently, Montbrun et al. reported the use of an OSATS to assess the technical competence of colorectal surgery residents (de Montbrun, Roberts et al. 2013). Combining bench, virtual reality, and cadaveric models, the OSATS was able to effectively discriminate between 10 graduating colorectal residents and 10 graduating general surgery fellows. Further study is required to assess the suitability of an OSATS in the setting of CBME - however, there is sufficient evidence to suggest that an OSATS would be a suitable tool for the in-training assessment of resident performance of technical procedures.

2.4.7 CONCLUSION

In the series of white papers released by the RCPSC, the college recommends the development of competency-based assessment tools that produce reliable and valid scores, with the establishment of a repository of tools and teaching strategies that can be adapted to different specialties. One of the major goals of this thesis will be to develop multiple assessment tools, using multiple examiners across a breadth of competencies, and used to determine the competence of residents in a CBME program. Certainly, this literature review provides evidence to support the use of OSCE’s to assess performance of clinical skills, and the application and understanding of the intrinsic roles, as well as the use of an OSATS to assess performance of technical procedures.

2.5 SUMMARY

It is clear that the transition from traditional time-based postgraduate medical training to CBME is in its initial phases. While much has been written about CBME, especially with regards its potential advantages and disadvantages, there is little in the way of objective data to support this new model of training. The introduction of CBME as the training format within the division of orthopedics at the University of Toronto offers a unique opportunity to examine some of the inherent concepts of CBME.
At the University of Toronto, a great deal of work has focused on curriculum development and program implementation - both faculty and residents are well acquainted with the model. Many issues remain, such as the increased costs related to CBME, the need for increased faculty involvement, as well as the logistics of how to manage a program that is learner-centred. However, the critical next step will be the development and evaluation of in-training assessment programs that can be used to determine competence of orthopedic residents. At this time, the majority of rotations in the division of orthopedics use oral examinations (typically administered by the faculty supervisor), short answer written examinations, and MCQs – assessment tools that suffer the limitations discussed in the literature review. While this remains an improvement over traditional training programs that did not use any form of in-training assessment, the increased use of valid and objective assessment tools to assess residency competency is critical to the implementation of a true outcomes-based program. These assessment tools must not only demonstrate evidence of validity, but be feasible and acceptable, and cover the breadth of competencies in surgical training: clinical skills, the Intrinsic CanMEDS Roles, and performance of technical procedures.

A major issue when considering this thesis surrounded the choice of assessment tools - whether to focus on WBA, to use simulation, or a combination of the two. Ultimately, tools such as OSCEs and OSATS were chosen for a number of reasons. Firstly, while there is good evidence for the use of WBA to provide formative feedback, CBME demands summative assessment to determine a minimal level of competence. Secondly, there is evidence that many WBAs have low inter- and intra-rater reliability, and are often used inappropriately trainees. Finally, there would be many logistical challenges in ensuring appropriate sampling of a rotation’s competencies using WBAs. For these reasons, the studies in this thesis will use assessment tools such as OSATS and OSCEs.

While CBME offers many advantages, a potentially poor assumption is the premise that all residents can become competent after a single rotation of subspecialty training. The CBME program at the University of Toronto dictates that residents must be deemed competent at each of 21 modules before moving to the next. However, a strict
competency-based approach ignores models of expertise and clinical reasoning, which generally theorise that most learners gradually progress towards competency over time. This will be a second important focus of research for this thesis; can residents achieve a minimal level of competence after a rotation within a CMBE model of training, or is there a more gradual transition to a minimal level of competence that occurs over time, related to experience and duration of training? The findings of this thesis may have wide-ranging implications for the future implementation of CBME.
Chapter 3
Aims and Hypotheses

AIMS

1. The overarching problem faced by CBME is inadequate assessment. At this time, educators are being asked to make competency decisions, often inappropriately. The first aim of this thesis is to develop a program of performance-based, in-training assessment tools that demonstrate evidence of validity, and are feasible and acceptable to both residents and faculty. These assessments will cover a breadth of competencies, including clinical skills, the Intrinsic CanMEDS Roles, and as well as the performance of technical procedures. Following this, this thesis will evaluate the assessment tools, using psychometric theory.

2. The second aim of this thesis will be to use the findings of these assessment tools to examine the premise that all residents can achieve a minimal level of competency after a subspecialty rotation. Using year of training as a surrogate of experience, a core principle of CBME will be examined, by determining if it is possible for both junior and senior residents to achieve competency after a rotation, or if there is an overall effect of experience.

RESEARCH QUESTIONS

1. What assessment tools can be developed for use as in-training assessment of resident competency, in the setting of CBME that demonstrate evidence of validity, and are both feasible and acceptable?

2. What is the effect of clinical experience in terms of time on the performance of orthopedic tasks - does the effect of clinical experience or duration of training have more of an impact than the training itself within a CBME model?
HYPOTHESES

Aim / Research Question 1

1. That an OSCE is a reliable method of assessing clinical skills after a sports medicine module, with evidence of validity

2. That the modified Angoff method is a credible and acceptable method of standard setting, able to set different standards for junior and senior residents in the setting of CBME

3. That an OSCE examination, designed to assess multiple intrinsic CanMEDS roles, would have sufficient reliability and validity to distinguish between different years of post-graduate training in orthopedic residents

4. That an Objective Structured Assessment of Technical skill (OSATS), using dry models, would be a reliable method of assessing residents’ ability to perform sports medicine procedures, and show evidence of validity

5. That a simulation-based assessment of resident performance of Entrustable Professional Activities would be reliable and show evidence of validity

Aim / Research Question 2

6. That with intensive training to a set curriculum, junior and senior residents would be able to demonstrate clinical skills to an equivalent level

7. That, after training in a CBME module, junior and senior residents would be able to perform simulated technical procedures to a similar level.

8. That there would there be evidence of important differences between junior and senior residents when performing simulated Entrustable Professional Activities
Chapter 4

Competency-Based Medical Education: Can Both Junior and Senior Residents Achieve Competence After a Sports Medicine Training Module?

T Dwyer, S Wright, M Kulasegaram, J Theodoropoulos, J Chahal, D Wasserstein, C Ringsted, Brian Hodges, D Ogilvie-Harris.
PMID: 26632001

In this study, I use an end of rotation OSCE to examine residents’ clinical skills after a sports medicine rotation, skills including history taking, examination, differential diagnosis, image interpretation, consent, clinical decision-making, and surgical techniques. This study demonstrates a significant difference between junior and senior residents in performance, with regards the total checklist scores and overall global rating scale. Despite all residents having been exposed to an identical curriculum and teaching program, and similar clinical training, significant differences in performance were also seen for each of the domains of knowledge (history-taking, consent, differential diagnosis etc.).
ABSTRACT

Background

Competency-based medical education (CBME) as a resident training format will move postgraduate training away from time-based training, to a model based upon observable outcomes. The purpose of this study was to determine whether junior and senior residents could demonstrate clinical skills to a similar level, after a sports medicine rotation. We hypothesized that, with intensive training to a set curriculum, junior and senior residents would be able to demonstrate clinical skills to an equivalent level.

Method

All residents undertaking a three-month sports medicine rotation had to pass an Objective Structured Clinical Examination (OSCE). The stations tested the fundamentals of history taking, examination, image interpretation, differential diagnosis, consent, and clinical decision-making. Stations were marked with a binary station-specific checklist and an overall global rating scale (GRS) (1=novice, 2=advanced beginner, 3=competent, 4= proficient, 5=expert). A GRS was also given for each domain of knowledge.

Results

Over 18 months, 39 residents (21 junior, 18 senior) and six fellows (45 participants) sat the examination. With regards to junior and senior residents, analysis using a two-tail t test demonstrated a significant difference in both total checklist score (%) (56.15 (SD 10.99) versus 71.87 (SD 8.94)) and overall global rating scale (2.44 (SD 0.55) versus 3.79 (0.49)) (p<0.01). There was also a significant difference between junior and senior residents for each knowledge domain (history p<0.05, remainder p<0.01). No significant difference was seen between the performance of senior residents and fellows.
Conclusion

Despite intensive teaching within a CBME model, junior residents were not able to demonstrate knowledge as well as senior residents, suggesting that overall experience is a critically important for achieving competency as measured by the OSCE.

INTRODUCTION

The division of orthopedics at our institution implemented a Competency-Based Medical Education (CBME) model of training on selected residents since 2009 (Alman, Ferguson et al. 2013, Ferguson, Kraemer et al. 2013). As of 2013, all first year residents were enrolled in a CBME program consisting of 21 rotations; residents must achieve a minimum level of competence in one before progressing to the next. In the event of a resident not meeting a predetermined standard, the rotation is prolonged or repeated until competence is achieved. The implementation of CBME as the resident training format has moved training from a time-based model, to a model based upon observable outcomes, creating a need for frequent, summative assessment (Pugh, Touchie et al. 2014).

One rotation is sports medicine, after which residents are expected to demonstrate competence in the management of patients with acute and chronic injuries of the knee, shoulder, hip, ankle and elbow. Residents undertake this rotation twice during their training, once as a junior (postgraduate year (PGY) 1 – 3) and once as a senior (PGY 4&5). A previous study identified that an Objective Structured Clinical Examination (OSCE) demonstrated evidence of validity and reliability when used to assess clinical skills in orthopedic residents (Dwyer, Theodoropoulos et al. 2013). The results of that study seemed to suggest that an increasing level of experience correlated to improved performance, more so than exposure to a recent sports medicine rotation. We hoped to investigate this issue further by comparing the performance of junior residents with those of senior residents after a sports rotation within a CBME model of training.
The purpose of this study was to determine whether junior and senior residents could demonstrate clinical skills to a similar level, after a single sports medicine rotation. We hypothesized, with intensive training to a set curriculum, that junior and senior residents would be able to demonstrate clinical skills to an equivalent level.

MATERIALS AND METHODS

Teaching sessions based upon a designated textbook (Kibler 2009) and curriculum were held three times a week. At the end of each rotation, residents undertook a six station OSCE assessing the fundamentals of history taking, examination, image interpretation, differential diagnosis, consent, and clinical decision-making. An established and previously described methodology (Dwyer, Theodoropoulos et al. 2013) was used to create the OSCE, mapped to the set curriculum, and focused on the CanMEDS Role of Medical Expert (Frank 2005).

Computer-based stations were created, with clinical photographs, imaging and intraoperative photographs displayed, in accordance with each scenario. At each station, faculty asked a series of predetermined questions – at three of the six stations, residents took a clinical history with faculty acting as a standardized patient. Approximately 50% of the examiners had been involved in teaching of the residents, as there are fewer residents than faculty. None of the junior residents had undertaken a prior sports rotation, while all of the senior residents had previously undertaken a sports rotation earlier, and a different sports OSCE in their previous rotation (Dwyer, Theodoropoulos et al. 2013). Sports medicine fellows at the beginning of their fellowship year were invited to participate – the majority had trained at other institutions, with a variety of exposure to sports medicine; all had recently passed the orthopedic certifying examination.

Residents rotated from station to station, marked by faculty using a station specific checklist and an overall global rating scale (GRS) – the GRS was a 5-point scale matched to the Dreyfus model of skill acquisition (1=novice, 2=advanced beginner, 3=competent, 4= proficient, 5=expert) (Dreyfus HL 2005). This GRS was also used for each clinical skill tested in each station. Examiners were instructed to assign a rating
of competent if residents performed at the passing level of the orthopedic certifying exam level, regardless of the postgraduate year of the trainee. While postgraduate year was unknown to examiners in approximately 50% of cases, it was not possible to blind examiners from the year of training for all residents. All examiners were experienced at rating residents within the OSCE format.

**Ethical Considerations**

Approval for this study was obtained from the institutional research ethics board.

**Source of Funding**

No external funding was used in this study.

**Statistical Analysis**

All data (checklists, global ratings) were de-identified, entered into an Excel spreadsheet and analyzed with the use of SPSS version 19 (Armonk, New York). Station metrics were calculated and analyzed based upon the recommendations outlined by AMEE guide no. 49 (Pell, Fuller et al. 2010). Reliability was calculated using Cronbach’s Alpha both for the total checklist scores (converted to a percentage), and for the total GRS (sum of the global ratings for each station and converted to a percentage). The contribution of each station to the overall reliability was observed using Cronbach’s Alpha if Item Deleted. Pearson’s correlation was sought for the relationship between the checklist scores and for the GRS for each station. ANOVAs were performed to identify any significant differences between examinee groups (junior, senior, fellows). Gabriel’s post hoc tests were used to identify significant differences between groups. In order to assess for the affect of familiarity on rating of performance, a univariate analysis was used to compare examiners familiar with residents to those examiners who were not for each station.

**RESULTS**

Over an 18-month period, 21 junior (PGY1, 2 &3), 18 senior (PGY4 & 5), and six sports fellows undertook the OSCE at the end of their rotation (45 participants).
Cronbach’s Alpha of the six stations was 0.89 for the total checklist and 0.87 for the overall global ratings. Cronbach’s Alpha if Station Removed demonstrated that no station removal increased the overall reliability of the OSCE, indicating that each station was making a positive contribution toward the overall reliability (Table 1). A good to high Pearson’s correlation was identified between the checklist scores and the single overall global ratings (0.65 – 0.89 for all stations), indicating that examiners were consistent in assigning low global ratings for performances resulting in low checklist scores, and vice versa.

<table>
<thead>
<tr>
<th>Station</th>
<th>Mean Percentage of Checklist Items Correct (SD)</th>
<th>Cronbach's Alpha if Station Removed</th>
<th>Mean Global (converted to percentage) (SD)</th>
<th>Cronbach’s Alpha if Station Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Patellofemoral dislocation</td>
<td>0.69 (0.13)</td>
<td>0.88</td>
<td>0.67 (0.17)</td>
<td>0.82</td>
</tr>
<tr>
<td>2. SLAP tear of the shoulder</td>
<td>0.65 (0.1)</td>
<td>0.89</td>
<td>0.64 (0.13)</td>
<td>0.87</td>
</tr>
<tr>
<td>3. Lateral ankle instability</td>
<td>0.57 (0.12)</td>
<td>0.88</td>
<td>0.62 (0.20)</td>
<td>0.84</td>
</tr>
<tr>
<td>4. Posterior cruciate ligament injury of the knee</td>
<td>0.69 (0.13)</td>
<td>0.87</td>
<td>0.70 (0.24)</td>
<td>0.84</td>
</tr>
<tr>
<td>5. Anterior Shoulder Dislocation</td>
<td>0.66 (0.13)</td>
<td>0.87</td>
<td>0.62 (0.18)</td>
<td>0.81</td>
</tr>
<tr>
<td>6. Elbow Osteochondritis Dissecans</td>
<td>0.56 (0.14)</td>
<td>0.89</td>
<td>0.53 (0.17)</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 1. Station metrics and Cronbach’s Alpha if Station Removed for checklist scores and overall global rating for each of the stations. SD – standard deviation, SLAP – superior labrum anterior posterior.
Figure 1 displays the mean total checklist scores and overall global ratings for each of the three groups. ANOVA testing demonstrated that there was a significant difference between groups (p<0.001) for both. A significant difference was also seen for all stations by checklist percent correct (p=0.001 for stations 1&2, p<0.001 for the remainder) and by overall GRS (p<0.001) except for station 2 (SLAP tear of the shoulder) (p=0.082).

![Figure 1](image.png)

**Figure 1.** Box plots for (A) overall global rating, and (B) total checklist score. A significant difference was seen between both junior and senior residents, and junior residents and fellows for both (p<0.01). No difference was seen between senior residents and fellows for either.

With regards to junior and senior residents, analysis using a two-tail t-test demonstrated a significant difference in both total checklist score (%) (56.15 (SD 10.99) versus 71.87 (SD 8.94)) (effect size (Cohen’s d) 1.57) and overall GRS (2.44 (SD 0.55) versus 3.79 (0.49)) (effect size 2.59) (p<0.01). Gabriel's post hoc analysis also demonstrated a significant difference between junior and senior residents for the total checklist scores (p<0.01), for the overall global rating on every station (p<0.01), and for each of the domains of clinical skill (p<0.01, history-taking (p<0.05)) (Figure 2).
Figure 2. Box plots for each of the domains of clinical skills (overall global rating scale). A significant difference was seen between junior and senior residents for every skill for each of the domains of clinical skill (all p<0.01 except history-taking (p<0.05)). No difference was seen between senior residents and fellows.
With regards to performance of senior residents and fellows, there was no significant difference for the overall checklist scores, the overall GRS, or on any station or domain of clinical skill. Using the mean checklist score and standard deviation of the fellows and the senior residents, a post hoc power analysis demonstrated that the observed power was 27%.

To determine if there was variations in reliability by group, a split Cronbach's alpha for the checklist scores by level of training was conducted, and was 0.8 for the juniors, 0.7 for the seniors, and 0.8 for the fellows. Variance components analysis showed that 64% of the variance was due to level of training, with differences between groups bigger than differences within groups. With regards to the effect of examiner familiarity with residents, subgroup analysis demonstrated that for stations 2-6 no effect was seen, regardless of familiarity status, on the total checklist score or global rating score. For station 1, there was a medium effect size for the global rating score only in favour of familiarity (Cohen's d=0.6, p=0.04).

In order to determine whether each resident passed the exam, a non-compensatory method was applied, whereby an overall global rating of competent had to be obtained at a minimum of 4/6 stations. Using this method, 8/21 (38%) junior residents, 18/18 (100%) senior residents, and 5/6 (83%) fellows passed the end of rotation exam.

**DISCUSSION**

The results of this study demonstrate that, despite training in a CBME model, junior residents were not able to demonstrate clinical skills as well as senior residents, with only 38% of junior residents passing this exam. These findings suggest that year of training, and by default overall clinical experience, is an important factor in attainment of competence. However, it may be that experience in conjunction with repeating the rotation twice is critical for the attainment of competence.

Many postgraduate programs are moving towards a CBME model (Cunningham, Kates et al. 2014), for potential benefits including improved training, the ability to identify residents requiring remediation (Lacasse, Theoret et al. 2014), and to
potentially shorten training. However, many of these proposed benefits have yet to be objectively demonstrated. A comparison of competency-based and traditional training in neurosurgery demonstrated that the time to perform complete procedures was reduced by 6 to 9 months with competency-based training (Long 2000). In critical care residents, it was shown that with an intensive, simulation-based education intervention, first-year residents outperformed traditionally trained third-year residents on a clinical skills assessment (Singer, Corbridge et al. 2013). Interestingly, both of these studies concentrate on the acquisition of technical skill, rather than the clinical skills examined in the current study.

A proposed advantage of CBME is the potential to shorten the length of residency to the amount of time taken for a resident to complete various competencies – a proposal that has both supporters (ten Cate and Scheele 2007) and opponents (Brooks 2009). A key feature of CBME is the ability to allow learners to progress at their own rate (Iobst, Sherbino et al. 2010) - some residents can achieve competency status earlier than required, while others require increased educational time (Ebert and Fox 2014). Reduced time in our residency program at this time has applied to a few residents, in whom frequent, summative assessment has allowed rapid advancement (Holmboe, Sherbino et al. 2010) – some of these residents were participants in this study. In the case of other residents, increased time in certain rotations has been beneficial.

This study found that junior residents failed the sports OSCE at a much higher rate than senior residents. We believe that models of expertise both predict and support this finding. Benner et al. studied theories of experiential (trial and error) learning in the field of nursing – nurses typically took a year to progress from a novice to an advanced beginner, then typically another 1.5 to 2 years to become competent (Benner 1996). The competent practitioner was distinguished from the advanced beginner by increased organizational ability, performing well in familiar situations. In line with this research, we believe that many orthopedic residents require time to acquire and organize knowledge into a logical form. The ACGME milestones for orthopaedic surgery support these theories (Sterns 2003), whereby residents are expected to progress through basic history-taking and examination (Level 1), take a
focused history and interpret basic imaging (Level 2), before progressing to recognize concomitant injuries, and being able to order and interpret advanced imaging. We believe that rather than expecting junior residents to become competent, it is important to ensure progression through milestones, and focus on identifying the resident in difficulty.

In recognition of this issue, our curriculum map emphasizes different objectives for the junior and senior sports rotations. One option would be to have different examinations for the junior and senior residents, however we continue to use a single OSCE for two reasons. Firstly, faculty in CBME programs risk being overwhelmed by the burden of assessment – at this time we run multiple examinations in the majority of our 21 rotations. Secondly, we believe that while junior residents should not be expected to become competent, they should be exposed to the minimal standards expected of a competent orthopedic surgeon early in their training. However, to avoid a potential situation whereby junior residents achieve the objectives that have been set for them but fail the exam, standard setting methods such as the modified Angoff method (Downing, Tekian et al. 2006) can be used to set different pass / fail standards for junior and senior residents. This is an area of research that we have addressed in a separate study.

In our curriculum, residents completed this rotation twice, emphasizing the importance of sports medicine within orthopedics (other rotations that are repeated include trauma and joint replacement). Whether or not junior residents would achieve a minimal level of competency if they were to undertake the rotation for six rather than three months is currently unknown. Studying this would require having two groups of residents – one group who sit two three-month rotations, and a second who undertake a single six-month rotation. This may well be important, as many of the other rotations are only three months long i.e. foot and ankle and spine. Further research is required to determine if a minimal level of competence could be achieved in three months in these rotations.

In the ACGME Orthopedic Milestones Project, sports medicine topics include anterior cruciate ligament, meniscal tear, and rotator cuff injury (Sterns 2003). While none of
these topics were examined in this OSCE, each one is part of our curriculum examined in a previous OSCE study (Dwyer, Theodoropoulos et al. 2013). Examples of topics covered in this OSCE include the management of patellofemoral joint dislocation, SLAP tears, as well as ankle and shoulder instability, clearly relevant to orthopedic training. It seems unlikely that the ACGME Orthopedic Milestones Project document is designed to act as a comprehensive orthopedic curriculum - rather it documents how the milestones concept should be applied to residency training.

A potential issue in our methodology was that the PGY level of some residents was known to examiners, raising the potential for bias. Subgroup analysis demonstrated that resident familiarity had an effect on one of the six stations, patellofemoral dislocation. Why there would be an effect of familiarity on one station only is unknown, and may be a chance finding. However, it is important to acknowledge that true objectivity is very difficult to obtain in most assessments, as the assessment of competency always requires the exercise of judgment by another competent, proficient or expert practitioner (Brooks 2009). An attempt to minimize these subjective effects is made via processes of standardization, by reaching consent regarding what is being evaluated, and the standards expected (Ginsburg, McIlroy et al. 2010). While high-stakes examinations avoid pairing examiners with familiar candidates, this is often impossible in low- or medium-stakes examinations (such as this one) due to cost, examiner availability and scheduling difficulties (Stroud, Herold et al. 2011). Our experienced examiners were asked to disregard the PGY level of residents - options to exclude the possibility of bias would be to use faculty from a separate institution, an option not typically feasible in the setting of regular assessment, or to videotape encounters. However, it does not seem likely that examiner familiarity resulting in bias would favour senior residents over junior residents, and thus affect the major findings of this study.

We are unable to comment on inter-observer reliability, with only one examiner at each station. However, other OSCE studies in other fields have demonstrated relatively high inter-rater reliability (Brown, Manogue et al. 1999). In orthopedics, Chen et al. videotaped an interaction between orthopaedic staff, fellows, and senior
residents, using a standardized patient with a femoral neck fracture - using a dichotomous pass/fail rating, the inter-rater reliability was >0.8 (Chen, Lee et al. 2013). Mohtadi et al. studied inter-rater reliability in a certifying examination for sports medicine diplomas; using two observers at 4/20 stations, the reliability co-efficient ranged from 0.85 to 0.99 (Mohtadi, Harasym et al. 1995).

Aggarwal and Darzi wrote that they believed expertise rather than experience should underlie competency-based practice (Aggarwal and Darzi 2006). At our institution we apply the CBME model, with teaching to a set curriculum, supported by regular, objective assessment. Despite this, this study provides evidence that experience and exposure are important components of the acquisition of clinical skill. The type of experience obtained by residents over the years in other rotations is hard to quantify, but would appear to be an important effect, and one that underpinned the traditional time-based models of surgical training for decades.

Limitations

The definition of an OSCE is variable (Harden and Gleeson 1979, Cohen, Reznick et al. 1990, Newble 2004) - stations may or may not involve a standardized patient (Hodges, Regehr et al. 1999). In three stations of this OSCE, residents were expected to take a history, with the examiner acting as the patient. Not using standardized patients (SPs) significantly reduces the costs of holding these examinations. It would be an option to use real patients rather than SPs – however, SPs have become the standard in OSCEs for reasons including availability, their ability to reliably present the same information over and over, and avoids potential mistreatment of real patients (exposure to inappropriate comments, poor examination techniques) (Barrows 1993).

The authors acknowledge that this study was underpowered to detect any significant difference between the fellows and senior residents – however our orthopedic training program is very large, and it would be difficult to increase numbers. We are unable to comment on inter-observer reliability, and we cannot say whether junior residents would become competent if exposed to two consecutive rotations. Finally, as discussed, when comparing outcomes based upon year of training, the potential for
bias must be considered. While an estimated 50% of examiners were unaware of the year of training of the resident, with evidence that the effect of familiarity was limited, it would require the use of examiners from other institutions or videotaping to completely exclude the risk of bias.

CONCLUSION

Despite intensive teaching within a CBME model, junior residents were not able to demonstrate clinical skills as well as senior residents, suggesting that overall experience is critically important for achieving competency as measured by the OSCE.
Chapter 5
How to set the bar in Competency-Based Medical Education: Standard setting using an Objective Structured Clinical Examination (OSCE)

T Dwyer, S Wright, KM Kulasegaram, J Theodoropoulos, J Chahal, D Wasserstein, C Ringsted, B Hodges, D Ogilvie-Harris
PMID: 26727954

The previous chapter demonstrated that, despite intensive teaching within a CBME model, junior residents were not able to demonstrate clinical skills as well as senior residents after a sports medicine rotation. Over 60% of junior residents failed the sports medicine OSCE – as a result, in this chapter I study a critical aspect of test development: standard setting. Both examinee-centred and test-centred methods are compared, enabling a discussion of how standards for OSCEs can be set in the context of CBME.
ABSTRACT

Background

The goal of the Objective Structured Clinical Examination (OSCE) in Competency-based Medical Education (CBME) is to establish a minimal level of competence. The purpose of this study was to 1) to determine the credibility and acceptability of the modified Angoff method of standard setting in the setting of CBME, using the Borderline Group (BG) method and the Borderline Regression (BLR) method as a reference standard; 2) to determine if it is feasible to set different standards for junior and senior residents, and 3) to determine the desired characteristics of the judges applying the modified Angoff method.

Methods

The results of a previous OSCE study (21 junior residents, 18 senior residents, and six fellows) were used. Three groups of judges performed the modified Angoff method for both junior and senior residents: 1) sports medicine surgeons, 2) non-sports medicine orthopedic surgeons, and 3) sports fellows. Judges defined a borderline resident as a resident performing at a level between competent and a novice at each station. For each checklist item, the judges answered yes or no for “will the borderline / advanced beginner examinee respond correctly to this item?” The pass mark was calculated by averaging the scores. This pass mark was compared to that created using both the BG and the BLR methods.

Results

A paired t-test showed that all examiner groups expected senior residents to get significantly higher percentage of checklist items correct compared to junior residents (all stations p<0.001). There were no significant differences due to judge type. For senior residents, there were no significant differences between the cut scores determined by the modified Angoff method and the BG / BLR method. For junior residents, the cut scores determined by the modified Angoff method were lower than the cut scores determined by the BG / BLR Method (all p<0.01).
Conclusion

The results of this study show that the modified Angoff method is an acceptable method of setting different pass marks for senior and junior residents. The use of this method enables both senior and junior residents to sit the same OSCE, preferable in the regular assessment environment of CBME.

Background

One of the key components of competency-based medical education (CBME) is regular assessment. At our institution, which has been trialing CBME in the postgraduate setting since 2009, the division of orthopedics has been using an end of rotation Objective Structured Clinical Examination (OSCE) to determine if residents are competent to progress to the next rotation - the goal of this in-training OSCE is to establish a minimal level of competence (Cusimano 1996).

Under the CBME curriculum at our university, residents undertake the sports medicine rotation twice, once as a junior resident (postgraduate year (PGY) 1-3), and once as a senior resident (PGY 4&5). While the curriculum map details different goals for junior and senior residents, for issues of feasibility all residents have been sitting the same OSCE. It was also believed that exposure of junior residents to the expected standard of clinical performance was important.

A previous study examining the use of an OSCE after a three months sports medicine rotation in a CBME program studied 45 participants (21 junior, 18 senior, six fellows), who undertook a six station OSCE over an 18-month period (unpublished data). The results of this OSCE (reliability >0.8), which tested the application of clinical skills, demonstrated a significant difference between junior and senior residents for the overall global ratings, total checklist scores, as well as for the global ratings / checklist scores for each station. Using a non-compensatory method (residents had to be deemed competent in 4/6 stations), only 8/21 (38%) of junior residents passed the exam, while 18/18 (100%) of senior residents passed.
Clearly, a pass rate of 38% for the junior residents appears unacceptable. Inherent in the two rotations for sports medicine is a belief that junior residents cannot become competent at sports medicine in a single rotation. Furthermore, different objectives had been set for the junior residents in the curriculum. In order to continue using a single OSCE, it was felt that a criterion-referenced standard setting method could be applied, whereby cut-scores would be set on the basis of what the resident should know, most suitable in the setting of CBME (Turnbull 1989, Norcini 1999, Norcini 2003, Schoonheim-Klein, Muijtjens et al. 2009).

Medical schools commonly use an examinee-centered standard setting method such as the mean Borderline Group (BG) method (Kaufman, Mann et al. 2000, Kilminster and Roberts 2004, Boursicot, Roberts et al. 2007). This method involves examiners in each station giving a global rating of each student’s overall performance independent of the checklist mark - the mean of scores for all candidates identified as borderline becomes the pass mark for the station, with the pass mark for the whole OSCE calculated by averaging each station’s borderline score (Wilkinson, Newble et al. 2001, Newble 2004). An alternative method used in the setting of small numbers of candidates (such as postgraduate orthopedic training) is the Borderline Regression (BLR) method, which predicts total OSCE scores from global ratings using linear regression, calculating the pass mark by substituting the score of borderline candidates into the regression equation for each OSCE (Wood, Humphrey-Murto et al. 2006, Schoonheim-Klein, Muijtjens et al. 2009, Hejri, Jalili et al. 2013). The downside of such methods in CBME is that the pass mark must be set after the exam has been undertaken.

An alternative method is a test-centered method whereby the pass mark is based on item or station characteristics, such as the modified Angoff standard setting method. In the modified Angoff methods, judges review each question after defining a borderline candidate, and decided whether the borderline examinee will respond correctly (Angoff 1971, Kaufman, Mann et al. 2000, Downing, Tekian et al. 2006, Schoonheim-Klein, Muijtjens et al. 2009). The main advantage of such a method is that the pass/fail standard can be reliably set before the OSCE is undertaken, useful in
the setting of CBME (Norcini 1992, Norcini 1997, Schoonheim-Klein, Muijtjens et al. 2009). Furthermore, a modified Angoff could potentially be used to set different pass marks for junior and senior residents, eliminating the need for two end of rotation OSCEs.

The purpose of this study was to 1) to determine the credibility and acceptability of the modified Angoff method of standard setting in the setting of CBME, using the BG and the BLR method as a reference standard; 2) to determine if it is feasible to set different standards for junior and senior residents, and 3) to determine the desired characteristics of the judges applying the modified Angoff method.

METHODS

The modified Angoff method.

In order to determine the most appropriate judges to establish competence in the setting of an orthopedic sports medicine rotation, three groups of judges were used: 1) sports medicine subspecialty surgeons, 2) non-sports medicine orthopedic surgeons, and 3) newly graduated orthopedic surgeons / orthopedic fellows. Six judges from each group participated, with this number based upon previous research (Brennan 1980, Williams, Klamen et al. 2003, Verheggen, Muijtjens et al. 2008). All judges, with the exception of the fellows, were members of faculty, experienced at both teaching and examining orthopedic residents – no judge had previous experience with standard setting methods.

Prior to the start of the OSCE, judges were asked to define an advanced beginner / borderline resident as a resident who is performing at a level between novice and competent at each station. For each checklist item for each station, the judges answered yes or no for “will the borderline examinee respond correctly to this item?” (Downing, Tekian et al. 2006, Schoonheim-Klein, Muijtjens et al. 2009). Items were then assigned as yes=1 and no=0, and the pass mark calculated by averaging the scores (Schoonheim-Klein, Muijtjens et al. 2009). Judges were asked to perform the modified Angoff method while separately imaging a borderline senior resident in PGY
4&5, and a borderline junior resident in PGY 1-3. The lead author discussed the modified Angoff method individually with each judge, and participated in the evaluation of the first 10 checklist questions on random stations.

**The Borderline Group Method**

After the conclusion of the study using an OSCE to examine the performance of residents after a sports medicine rotation, the residents who were rated advanced beginner for each station were called the borderline group for that station. The mean checklist score of this group was calculated and used as the station pass mark (Kaufman, Mann et al. 2000, Wilkinson, Newble et al. 2001, Newble 2004, Boursicot, Roberts et al. 2007).

**The Borderline Regression (BLR) method**

The Borderline Regression (BLR) method was used to calculate pass marks based on the results of a regression analysis, using the global station mark as the independent variable and the station checklist mark as the dependent variable. The regression line was calculated based on all of the interactions between examiners and students for a particular station. The pass mark was then the point at which the regression line crosses the borderline category. The overall pass mark using BLR method was calculated as the total of all of the station pass marks plus 1 Standard Error of Measurement (SEM) (roughly equivalent to 0.5*SD of the overall score). The SEM is calculated by taking the square root of (1-cronbach’s alpha for checklist total)* SD.

**Ethical Consideration**

Approval for this study was obtained from Women’s College Hospital Research Ethics Board, approval number 2012-0009-E.

**Statistical Analysis**

The mean cut-scores / passing scores, pass rates (%), confidence intervals, and SEMs were calculated for each of the methods, for each station and overall. Differences between pass/fail standards of the modified Angoff, the BG, and the BLR
methods were tested using a paired t-test, considering $p < 0.05$ as statistically significant (Kramer, Muijtjens et al. 2003). The inter-rater judge agreement for the Angoff method was calculated for each item, station and overall, using an intra-class correlation coefficient. The acceptability (satisfactoriness) or impact of each standard setting method on exam level pass/fail decisions was examined, based upon the passing rates (percentage of residents passing) of the junior residents by each of the methods (Downing, Lieska et al. 2003). Credibility was judged by the number of PGY4&5 residents / orthopedic fellows passing the OSCE determined by each of the methods (Kramer, Muijtjens et al. 2003).

RESULTS

Modified Angoff Method by examiner type

A total of six sports surgeons, six non-sports surgeons, and six fellows undertook the modified Angoff method for the OSCE. A paired samples t-test showed that all examiner groups (based upon the mean for all three examiner types) expected senior residents to get a significantly higher percentage of checklist items correct compared to junior residents (all stations $p<0.001$) (Table 2). While there was a tendency for non-sports surgeons to have the highest expectations for senior (and in some cases, junior) residents, there were no significant differences due to examiner type (all stations $p >0.05$). There was a high correlation (ICC) between judges for the modified Angoff method for both the junior residents (0.85) and for the senior residents (0.9).
Table 2. Modified Angoff Method. Displayed are the expected checklist percentage correct by examiner type and by resident group (junior – PGY1-3, senior PGY 4&5). For all three groups of judges (six in each group), a significant difference was seen both overall and for each station between junior and senior residents (p<0.001). No significant difference was seen between each group of judges. PGY – postgraduate year.

### Differences Between Standard Setting Methods

The pass marks established by the BG method, BLR method and modified Angoff methods are seen in Table 3 and 4. For senior residents, there were no significant differences between the pass marks determined by the modified Angoff method and the BG / BLR methods. For junior residents, there were significant differences for all stations and for the overall pass mark, with the pass marks determined by the Angoff

<table>
<thead>
<tr>
<th>Station</th>
<th>Residents</th>
<th>Sports Surgeon (95% CI)</th>
<th>Non Sports Surgeon (95% CI)</th>
<th>Fellow (95% CI)</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Senior</td>
<td>0.57 (0.4-0.74)</td>
<td>0.64 (0.48-0.8)</td>
<td>0.58 (0.49-0.67)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Junior</td>
<td>0.37 (0.23-0.51)</td>
<td>0.37 (0.28-0.47)</td>
<td>0.33 (0.26-0.39)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Senior</td>
<td>0.56 (0.37-0.75)</td>
<td>0.67 (0.44-0.9)</td>
<td>0.58 (0.47-0.69)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Junior</td>
<td>0.43 (0.26-0.6)</td>
<td>0.39 (0.25-0.52)</td>
<td>0.36 (0.26-0.46)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Senior</td>
<td>0.46 (0.31-0.61)</td>
<td>0.57 (0.36-0.78)</td>
<td>0.49 (0.35-0.64)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Junior</td>
<td>0.35 (0.21-0.49)</td>
<td>0.34 (0.21-0.47)</td>
<td>0.29 (0.22-0.37)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Senior</td>
<td>0.58 (0.27-0.76)</td>
<td>0.66 (0.42-0.89)</td>
<td>0.5 (0.37-0.63)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Junior</td>
<td>0.33 (0.18-0.49)</td>
<td>0.34 (0.21-0.49)</td>
<td>0.29 (0.24-0.35)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Senior</td>
<td>0.56 (0.41-0.71)</td>
<td>0.71 (0.51-0.91)</td>
<td>0.56 (0.41-0.72)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Junior</td>
<td>0.38 (0.21-0.54)</td>
<td>0.41 (0.21-0.61)</td>
<td>0.36 (0.23-0.5)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Senior</td>
<td>0.50 (0.35-0.65)</td>
<td>0.67 (0.49-0.84)</td>
<td>0.5 (0.31-0.69)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Junior</td>
<td>0.35 (0.21-0.5)</td>
<td>0.42 (0.27-0.57)</td>
<td>0.36 (0.24-0.48)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean Senior</td>
<td>0.53 (0.42-0.63)</td>
<td>0.65 (0.5-0.8)</td>
<td>0.54 (0.42-0.65)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Mean Junior</td>
<td>0.37 (0.27-0.47)</td>
<td>0.38 (0.27-0.49)</td>
<td>0.33 (0.26-0.41)</td>
<td></td>
</tr>
</tbody>
</table>
method were significantly lower than the pass marks determined by the BG / BLR Method (all p<0.01).

The number of residents who failed each station and failed overall using each standard setting method are seen in Table 5, with a comparison of each method seen in Figure 3. Using the BG and the BLR method, 6/21 (28.6%) of junior residents failed the OSCE; using the modified Angoff method, only 1/21 (4.8%) junior residents failed. Using the BG, the BLR method, and the modified Angoff method, no senior resident or fellow failed the OSCE.

<table>
<thead>
<tr>
<th>Station</th>
<th>Modified Angoff method (95% CI)</th>
<th>Borderline Groups Method (95% CI)</th>
<th>Borderline Regression Method (95% CI)</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6 (0.54-0.66)</td>
<td>0.53 (0.52-0.54)</td>
<td>0.52 (0.46-0.58)</td>
<td>n.s.</td>
</tr>
<tr>
<td>2</td>
<td>0.6 (0.52-0.68)</td>
<td>0.54 (0.53-0.56)</td>
<td>0.56 (0.51-0.61)</td>
<td>n.s.</td>
</tr>
<tr>
<td>3</td>
<td>0.51 (0.43-0.59)</td>
<td>0.48 (0.47-0.48)</td>
<td>0.47 (0.44-0.51)</td>
<td>n.s.</td>
</tr>
<tr>
<td>4</td>
<td>0.55 (0.46-0.65)</td>
<td>0.58 (0.49-0.66)</td>
<td>0.55 (0.46-0.64)</td>
<td>n.s.</td>
</tr>
<tr>
<td>5</td>
<td>0.61 (0.52-0.7)</td>
<td>0.54 (0.53-0.54)</td>
<td>0.56 (0.51-0.6)</td>
<td>n.s.</td>
</tr>
<tr>
<td>6</td>
<td>0.55 (0.46-0.64)</td>
<td>0.5 (0.48-0.51)</td>
<td>0.51 (0.47-0.54)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Total</td>
<td>0.57 (0.49)-(0.65)</td>
<td>0.53 (0.5-0.55)</td>
<td>0.53 (0.43-0.62)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Total + 1SEM</td>
<td>0.60</td>
<td>0.56</td>
<td>0.56</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

**Table 3.** Pass marks using modified Angoff for the senior residents, Borderline Groups (BG) and Borderline Regression (BLR) method. No significant difference was seen on any station between the modified Angoff method and the BG / BLR methods for senior residents. n.s. – non significant.
### Table 4. Pass marks using modified Angoff for the junior residents, Borderline Groups (BG) and Borderline Regression (BLR) method. A significant difference was seen for the modified Angoff and the BG / BLR methods for junior residents for all stations and overall (all p<0.001).

<table>
<thead>
<tr>
<th>Station</th>
<th>Modified Angoff method (95% CI)</th>
<th>Borderline Groups Method (95% CI)</th>
<th>Borderline Regression Method (95% CI)</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.35 (0.31-0.4)</td>
<td>0.53 (0.52-0.54)</td>
<td>0.52 (0.46-0.58)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.39 (0.33-0.45)</td>
<td>0.54 (0.53-0.56)</td>
<td>0.56 (0.51-0.61)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>3</td>
<td>0.32 (0.27-0.37)</td>
<td>0.48 (0.47-0.48)</td>
<td>0.47 (0.44-0.51)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>4</td>
<td>0.32 (0.27-0.37)</td>
<td>0.58 (0.49-0.66)</td>
<td>0.55 (0.46-0.64)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>5</td>
<td>0.38 (0.31-0.46)</td>
<td>0.54 (0.53-0.54)</td>
<td>0.56 (0.51-0.6)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>6</td>
<td>0.38 (0.31-0.44)</td>
<td>0.5 (0.48-0.51)</td>
<td>0.51 (0.47-0.54)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>Total</td>
<td>0.36 (0.3-0.42)</td>
<td>0.53 (0.5-0.55)</td>
<td>0.53 (0.43-0.62)</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>Total + 1SEM</td>
<td><strong>0.39</strong></td>
<td><strong>0.56</strong></td>
<td><strong>0.56</strong></td>
<td>P&lt;0.001</td>
</tr>
</tbody>
</table>
Figure 3. Comparison of cut-scores using the modified Angoff, Borderline Group (BG), and Borderline Regression (BLR) methods for junior and senior residents for each station.

<table>
<thead>
<tr>
<th>Station</th>
<th>Modified Angoff method</th>
<th>Borderline Groups Method</th>
<th>Borderline Regression Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Junior (n=21)</td>
<td>Senior (n=18)</td>
<td>Fellow (n=6)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Overall</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5. Number of failures established by each standard setting method, for each station and overall.
DISCUSSION

To our knowledge, this is the first study to examine the use of standard setting methods in the setting of postgraduate CBME. The results of this study demonstrate that the modified Angoff method can be used to establish different pass marks for junior and senior residents in the setting of CBME. We have also demonstrated that the subspecialty training of judges does not significantly change pass marks.

One of the challenges of the CBME model is the organization of frequent, objective assessments, requiring considerable faculty involvement and resources. The curriculum at our institution has 21 different rotations – each resident is required to demonstrate a minimal level of competence in one rotation before progression to the next. The sports medicine rotation is one of the major rotations, which residents undertake both as a junior and as a senior, with the curriculum map detailing different objectives for each group. While previous research had demonstrated that the junior residents could not master clinical skills as well as the seniors, we would prefer to continue using a single OSCE in the interest of feasibility (Dwyer, Theodoropoulos et al. 2013). Using the modified Angoff method to set different pass marks for junior and senior residents allows us to do so. In this manner, we can also identify the one or two junior residents who are performing poorly compared to their peers.

The advantages of the modified Angoff method include its relative simplicity, as well as the fact that the pass/fail standard can be set before the OSCE is undertaken (Schoonheim-Klein, Muijtjens et al. 2009). Initially, the authors of this paper had been concerned that sports medicine specialists, who were involved in content review and exam writing for the sports rotation, would set the pass mark too high, especially for the junior residents. Interestingly, although there was no significant difference between the groups, there was a trend for the non-sports surgeons to demand more, especially from the senior residents. These results are encouraging; it seems logical to have those surgeons involved in the content development perform the standard setting procedure.
In 2000, Kaufman et al compared the Angoff method (different from the modified Angoff, whereby judges estimate ‘the probability that a ‘minimally acceptable’ or borderline candidate would answer an item correctly’) (Angoff 1971) and the BG method, and found that both provide reasonable and defensible results in the medical school setting (Kaufman, Mann et al. 2000). In contrast, Kramer et al. examined standard setting in postgraduate general practice training, identifying that the BLR method was more credible and acceptable than the modified Angoff method (Kramer, Muijtjens et al. 2003).

These conflicting findings may be the results of some known difficulties with the Angoff method. Verheggen et al. demonstrated considerable variation between judges, especially when judges had less expertise in certain item areas (Verheggen, Muijtjens et al. 2008). The authors recommended careful judge selection, and that judges should be capable of answering all items correctly. In the study by Kramer et al., 84 examiners were used, the majority of whom also performed the modified Angoff methods (Kramer, Muijtjens et al. 2003). In our study only six judges were used, all of who were involved in exam creation and acted as OSCE examiners. This may explain why the modified Angoff method was shown to be credible and acceptable in our setting.

In this study, in order to perform the BG method and the BLR method, we were able to use the results of 45 participants, who sat the OSCE over 18 months, However, as an OSCE becomes a common assessment method within our CBME program, waiting to generate sufficient results before perform a BG / BLR method is not acceptable. For this reason, the findings that the pass marks established by the modified Angoff method are acceptable and credible, and can be performed by the subspecialty judges are extremely significant, and are of potential value to similar sized CBME based residency programs.

The results of the OSCE study demonstrated that while senior residents were able to achieve a minimal level of competence, junior residents were not (research to be published in December 2015). The results of this study are not dissimilar to studies that use progress testing, whereby regular assessment throughout an academic
program are used to provide longitudinal evidence of the growth of student knowledge (Wrigley, van der Vleuten et al. 2012). The sports medicine OSCE was not designed as a progress test, but rather to determine whether both senior and junior residents could achieve a minimal level of competence. However, continued iterations of the OSCE will be used in a manner similar to progress testing, to ensure that all residents are performing as expected in comparison with their peer group.

One of the main limitations of this study was an inability to demonstrate the credibility of using the modified Angoff to establish a pass mark for the junior residents. Credibility for the use of the modified Angoff to set pass marks for senior residents was established by a comparison with the pass marks established using the BLR / BG methods. However, in the setting of the junior residents, the pass mark created was significantly lower than that set by the BLR / BG methods. While some credibility for this standard setting method was demonstrated by the finding that all three groups of judges set similar cut scores for junior residents, there was no alternative standard setting method that could be used for comparison.

Downing wrote that there is no single correct answer when comparing standard-setting methods (Downing 2003), with Norcini and Shea stating that issues of student fairness are the most important – the passing scores must be acceptable to students, faculty and administrators (Norcini 1997). Having nearly 30% of junior residents fail using the BG / BLR methods certainly appears unacceptable. The purpose of the end of rotation OSCE was to identify those residents not performing as well as their peer group, which the modified Angoff method appears to effectively do. Credibility can also be established by using a systematic approach, produced by qualified judges with a clear purpose (Norcini 1997), which was the case in the study. Methods should also be supported by a body of published research, be transparent, easy to implement and easy to explain – such methods justify the final result (Norcini 1992).

Other limitations of this study include the use of only six judges in each group to perform the modified Angoff method, despite evidence that increased number of judges improve the reliability of the modified Angoff – however there was a high correlation between judges for the modified Angoff, and the pass marks created for the
senior residents matched that set by the BG / BLR method. In this study, the credibility as opposed to the validity of standard setting methods was studied, with credibility established by comparing the pass/fail rates of different methods with a reference group that is expected to have a high pass rate (Kramer, Muijtjens et al. 2003). While these two terms could be used interchangeable, credibility is typically used in the standard setting literature and was thus used in this study (Kramer, Muijtjens et al. 2003). Finally, this study also uses the OSCE results of relatively few residents, especially in comparison to other studies that have used the results of medical students - however it would be difficult to increase these numbers in the setting of postgraduate orthopedic training without performing a multi-centred study.

CONCLUSION

The results of this study demonstrate that the modified Angoff method can be used to set acceptable and credible cut-scores for junior and senior residents sitting an end of rotation OSCE in the setting of postgraduate CBME. This allows the use of the modified Angoff method to establish separate junior and senior cut-scores before each OSCE, not just in sports medicine, but also in other rotations such as joint replacement and trauma.
Chapter 6
How to assess Communication, Professionalism, Collaboration and the other Intrinsic CanMEDS Roles in orthopaedic residents: Use of an Objective Structured Clinical Assessment (OSCE).


This chapter focuses on the assessment of the intrinsic (non medical expert) CanMEDS roles, presenting the results of a study that used an OSCE to assess the understanding and application of these roles in three groups of residents; PGY-0 (new residents), PGY-3 and PGY-5. Many faculty members postulated that junior residents would be able to manage difficult clinical scenarios as well as seniors. After the OSCE, the results of which demonstrated that senior residents outperformed junior residents, it seemed evident that for many residents, previous experience with difficult clinical situations in their training (cancelled patients, delivering bad news) had provided them with important skills useful in managing these types of scenarios.
ABSTRACT

Background

Assessing residents’ understanding and application of the six CanMEDS intrinsic roles (Communication, Professional, Manager, Collaborator, Health Advocate, Scholar) continues to be a challenge for postgraduate medical educators. We hypothesized that an OSCE examination, designed to assess multiple intrinsic CanMEDS roles would have sufficiently reliable and valid.

Methods

The OSCE was composed of 6 ten-minute stations - stations tested two intrinsic roles using case-based scenarios (with or without the use of standardized patients). Residents were evaluated using five-point performance rating scales, as well as an overall rating of performance by an orthopedic surgeon at each station. Concurrent validity was sought by correlation with In-Training Evaluation Reports (ITERs) from the last 12 months, and an Ordinal Ranking created by Program Directors (PDs)

Results

Twenty-five residents from PGY 0, 3 and 5 participated. The inter-station reliability for total test scores (percent) was 0.87, while reliability for each of the Communicator, Collaborator, Manager and Professional roles was > 0.8. Total test scores, individual station scores and individual CanMEDS role scores all showed a significant effect by PGY level. Analysis of the PD rankings of intrinsic roles demonstrated a high correlation with the OSCE role Scores. A correlation was seen between the Communicator roles on the ITER and the OSCE, while the ITER Medical Expert and Total Scores highly correlated with the Communicator, Manager and Professional OSCE scores.
Conclusions

An OSCE designed to assess the intrinsic CanMEDS roles proved to be sufficiently valid and reliable to be used regularly in an orthopedic residency program.

INTRODUCTION

The seven CanMEDS competencies (Medical Expert, Communicator, Collaborator, Manager, Health Advocate, Scholar and Professional) have been clearly outlined in the CanMEDS 2005 Physician Competency Framework, by the Royal College of Physicians and Surgeons of Canada (Frank 2005). A similar framework has been described by the Accreditation Council for Graduate Medical Education (ACGME), defining six core competencies (2007). Each of these frameworks describe the principal generic abilities of physicians in health care, and are an integral component of postgraduate education. However, despite the widespread popularity of the CanMEDS roles and other competency frameworks, as well as a mandate to both teach and assess these competencies, the best methods remain unknown (Jefferies, Simmons et al. 2011, Zuckerman, Holder et al. 2012).

Assessment options for the intrinsic roles (non Medical Expert) include ITERs (in-training evaluation reports), structured oral examinations, 360-degree assessments, and objective structure clinical examinations (OSCE) (Zuckerman, Holder et al. 2012). A survey of a wide variety of medical and surgical program directors in Canada identified that the ITER is the most commonly used method to evaluate the CanMEDS roles, despite its acknowledged subjective nature (Catton 1997, Jefferies, Simmons et al. 2007, Chou, Cole et al. 2008). Respondents reported dissatisfaction with current methods of evaluating the intrinsic roles, especially Manager and Health Advocate.

The OSCE is a term used to describe a variety of multi-station examinations, and is a format currently favored at orthopedic certifying examinations in Canada and other countries worldwide. Studies using OSCEs to assess the role of medical expert have demonstrated reliability and validity in post-graduate physician training (Cohen, Reznick et al. 1990, Hodges, Regehr et al. 1998, Schwartz, Witzke et al. 1998,

There is also some evidence that an OSCE can be used to assess other CanMEDS roles, including the application of evidence-base medicine or demonstration of teaching skills (Scholar) (Schol 2001, Fliegel, Frohna et al. 2002), as well as cultural awareness and the application of ethical principles (Professional) (Altshuler and Kachur 2001, Singer, Pellegrino et al. 2001). An OSCE has also been used to assess multiple CanMEDS competencies in other fields of postgraduate training, such as radiology and neonatology (Jefferies, Simmons et al. 2007, Probyn 2010).

To our knowledge, no research exists regarding methods of assessing the intrinsic roles in orthopedic postgraduate training. We hypothesized that an OSCE examination, designed to assess multiple intrinsic CanMEDS roles, would have sufficient reliability and validity to distinguish between different years of post-graduate training in orthopaedic residents.

**MATERIALS AND METHODS**

**Exam Development**

The Orthopedic Residency Program at the University of Toronto, in collaboration with the Postgraduate Medical Education department, designed an orthopedic OSCE to test the six intrinsic CanMEDS roles. A focus group of academic orthopedic specialists was assembled with the goal to create clinical scenarios evaluating selected CanMEDS principles for each of the roles. The focus group relied on the source document from The Royal College of Physician and Surgeons of Canada, in which each of the roles, and the key competencies of each, is clearly defined (Frank 2005).
The OSCE was one hour long and composed of 6 ten-minute stations. The roles of Communicator, Collaborator, Professional, Manager, Health Advocate, and Scholar were assessed; a deliberate attempt was made to avoid testing Medical Expert. The majority of the six case-based scenarios were designed to assess a primary and secondary role. Stations 2 – 4 used standardized patients (SP’s) (station 2: relative concerned regarding delay in surgery; station 3: grandmother of child with suspected non accidental injury; station 4: teenager being informed of osteosarcoma diagnosis), while station 5 utilized a standardized health professional (SHP) (operating room manager). Two stations did not have a SP or SHP (station 1: ethical approach to needlestick injuries; station 6: evidence based medicine in spinal surgery). Table 6 lists the roles, and associated key competencies tested in each station.
<table>
<thead>
<tr>
<th>Station</th>
<th>Roles</th>
<th>Key competencies tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Needlestick</td>
<td>1° Professional</td>
<td>• Bioethical principles&lt;br&gt;• Informed consent / confidentiality&lt;br&gt;• Commitment to professional standards</td>
</tr>
<tr>
<td></td>
<td>2° Manager</td>
<td>• Priority setting&lt;br&gt;• Time management</td>
</tr>
<tr>
<td>2. Trauma list</td>
<td>1° Manager</td>
<td>• Task prioritization and time management&lt;br&gt;• Leadership&lt;br&gt;• Appropriate use of resources in hospital</td>
</tr>
<tr>
<td></td>
<td>2° Communicator</td>
<td>• Effective listening&lt;br&gt;• Empathy&lt;br&gt;• Patient-centred approach to communication</td>
</tr>
<tr>
<td>3. Non accidental injury</td>
<td>1° Health Advocate</td>
<td>• Recognition of patient risk factors&lt;br&gt;• Modification of risk factors&lt;br&gt;• Patient safety&lt;br&gt;• Leadership&lt;br&gt;• Negotiation</td>
</tr>
<tr>
<td></td>
<td>2° Manager</td>
<td></td>
</tr>
<tr>
<td>4. Breaking bad news</td>
<td>1° Communicator</td>
<td>• Breaking bad news&lt;br&gt;• Addressing end of life issues&lt;br&gt;• Empathy&lt;br&gt;• Autonomy in decision making&lt;br&gt;• Bioethical principles</td>
</tr>
<tr>
<td></td>
<td>2° Professional</td>
<td></td>
</tr>
<tr>
<td>5. Interacting with OR team</td>
<td>1° Collaborator</td>
<td>• Conflict resolution&lt;br&gt;• Respect for members of the healthcare team&lt;br&gt;• Recognizing one’s own roles and limitations</td>
</tr>
<tr>
<td>6. Spinal evidence</td>
<td>1° Scholar</td>
<td>• Evidence based medicine&lt;br&gt;• Critical appraisal of evidence&lt;br&gt;• Translating knowledge into practice</td>
</tr>
<tr>
<td></td>
<td>2° Communicator</td>
<td>• Interactive process&lt;br&gt;• Efficiency and accuracy</td>
</tr>
</tbody>
</table>

**Table 6.** Breakdown of each individual station by primary and secondary roles, and the key competencies tested, taken from the CanMEDS 2005 Physician Competency Framework (Frank 2005). Communicator was also assessed in stations 1 and 3.
The CanMEDS OSCE development was facilitated by an exam blueprint and case development guides. A member of the focus group was assigned as lead to design each station, which was then reviewed by the entire focus group. Any discrepancies or ambiguities were addressed or removed. A number of 5-point performance rating scales were developed for each of the intrinsic roles. The ratings were anchored by descriptions of performance to be demonstrated by the residents for each role. An overall 5-point global rating was also assigned for each resident at the end of the station. The SP’s and SHP’s were selected from an established Standardized Patient Bank at the University of Toronto. For the OSCE, two SP’s / SHP’s were trained for each of the stations by an experienced SP trainer; each received a minimum of three hours training for each role. No SP or SHP assessment of performance was used.

**Study design**

Convenience sampling was used to recruit residents from specific postgraduate years (PGY) of training. This group involved incoming residents who had not yet begun residency training, assigned to the group PGY0. The second group was composed of PGY3 residents at the end of their year of training, with volunteers from the PGY5 residents also asked to sit the OSCE. These PGY5 residents had all recently passed their orthopedic certification examinations, and were used the “gold standard”.

Members of the orthopaedic faculty at each station evaluated residents independently. It was not possible to blind examiners from the year of training of the residents, as many of the residents were familiar to the staff surgeons. However, examiners were asked to disregard the year of training when making assessments. The OSCE was conducted in four one-hour sessions over the course of a single day. Each candidate signed a consent form permitting the use of exam results for research purpose. On completion of the exam, residents were invited to provide feedback using a five-point Likert scale. Summative and formative feedback was given to each resident at the end of the OSCE.

Concurrent validity as sought in two ways. Firstly, the ITERs from the preceding 12 months were obtained for the PGY3 and PGY5 residents, and the results on the six
intrinsic roles correlated with the OSCE total score and role scores. Secondly, the two Program Directors (PD) formed an ordinal ranking of the residents in PGY3 & 5, and rated each residents’ ability in each of the CanMEDS roles using a five point scale (1=needs significant improvement, 2=below expectations, 3=solid competent performance, 4=exceeds expectations, 5=superb). The overall ranking, and the rating for each role were also correlated to the total OSCE score and role scores.

Statistical Analysis

All data was de-identified, and residents assigned a study specific number. Raw scores from the individual station scores and role scores were entered into a spreadsheet, and analyzed using SPSS version 19 (Armonk, NY). All scores were converted into a percentage, with results reported as means +/- standard deviation. Reliability was established using the inter-station alpha co-efficient of reliability (Cronbach’s alpha) for each of the scoring tools. Scores from the different rating scales were evaluated using regression analysis. The effect of PGY on total test scores (%), overall ratings of performance, individual station scores and role scores were evaluated using one-way analysis of variance (ANOVA). A p value of < 0.05 was considered to be significant. Scheffe’s test was used for post hoc analysis to understand differences in scores between each possible pair of year of training. Correlation between total OSCE scores and role scores with ITER role scores and PD rankings was using Pearson’s correlation and Spearman’s Rho ($R^2$). Student T-test was used to compare the PD ratings of resident performance between PGY.

Ethics Approval

Approval for this study was obtained from the Research Ethics Board, University of Toronto. Each resident signed a consent form to permit the use of the OSCE results, and ITER results for research purposes.

RESULTS

Twenty-five residents from PGY0, 3 and 5 took part in the OSCE (Table 7). The roles of Communicator, Manager and Professional were assessed in multiple stations;
Collaborator, Health Advocate and Scholar were assessed in only one station each. The total test scores (converted to a percentage) and the mean overall rating of performance are seen in Table 8; ANOVA testing demonstrated a significant difference of the effect of PGY on both scores (p < 0.001). A significant difference was seen between PGY0 and PGY3 (p<0.05), PGY3 and PGY5 (p=0.001), and PGY0 and PGY5 (p<0.001).

<table>
<thead>
<tr>
<th>PGY</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGY 0</td>
<td>6</td>
</tr>
<tr>
<td>PGY 3</td>
<td>13</td>
</tr>
<tr>
<td>PGY 5</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
</tr>
</tbody>
</table>

**Table 7.** Resident demographics.

<table>
<thead>
<tr>
<th>Total Test Score</th>
<th>Grand Mean (SD)</th>
<th>PGY0 Mean (SD)</th>
<th>PGY3 Mean (SD)</th>
<th>PGY5 Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/100</td>
<td>75% (12.9)</td>
<td>62.4% (5)</td>
<td>73.4% (9)</td>
<td>91.1% (8.3)</td>
</tr>
<tr>
<td>/5 (Overall Rating of Performance)</td>
<td>3.55 (0.78)</td>
<td>2.75 (0.35)</td>
<td>3.45 (0.50)</td>
<td>4.56 (0.44)</td>
</tr>
</tbody>
</table>

**Table 8.** Total test scores and overall rating of performance.
The inter-station reliability for total test scores (percent) was 0.87, and the inter-station reliability for overall ratings of performance was 0.83. The internal consistency for four of the six role scores are seen in Table 9; internal consistency coefficients were not able to be computed for Scholar or Advocate roles, as only one rating scale was used for each of these Roles. The consistency for each of these four roles was very high (>0.80).

<table>
<thead>
<tr>
<th>Role</th>
<th>Alpha coefficients</th>
<th>Item numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communicator</td>
<td>.91</td>
<td>17 items in 5 of 6 stations</td>
</tr>
<tr>
<td>Collaborator</td>
<td>.96</td>
<td>3 items in 1 of 6 stations</td>
</tr>
<tr>
<td>Manager</td>
<td>.83</td>
<td>5 items in 3 of 6 stations</td>
</tr>
<tr>
<td>Professional</td>
<td>.84</td>
<td>3 items in 2 of 6 stations</td>
</tr>
</tbody>
</table>

*Table 9.* Internal consistency for the four Role scores with more than one rating scale.

The total test scores for the individual stations by PGY are displayed in Figure 4. The effect of PGY on the individual station scores was statistically significant (station 1,5&6 p<0.01; station 2&4 p<0.05) with the exception of station 3 (p=0.068). Post-hoc analysis demonstrated a significant difference in station scores between PGY5 and PGY0 / PGY3, except for Station 3. No significant difference was seen between scores for PGY 0 & 3, although a trend was seen for increased scores with PGY3.
Figure 4. Total station scores (% correct) by PGY for each of the stations. Each station showed a significant difference by PGY (p<0.05) except for station 3 (NAI / Non accidental injury) (p=0.068). Error bars represent standard error of the mean. PGY – postgraduate year.

The total test scores for each of the intrinsic roles by PGY are seen in Figure 5. ANOVA testing for the effect of PGY on each of the role scores was statistically significant (Communicator, Collaborator, Manager, Professional; p<0.001; Health Advocate, Scholar p<0.05). For each of the role scores, PGY0 and PGY3 were significantly different from PGY5, but not from each other.
Figure 5. Total test scores (% correct) for each of the CanMEDS roles by PGY. ANOVA testing showed significant differences for all roles (communicator, collaborator, manager, profession p<0.001; advocate, scholar p<0.05). PGY – postgraduate year.

Analysis of the PD ratings of intrinsic roles demonstrated a good correlation between these and the corresponding OSCE role scores (Table 10). ITERs were available for 12 months prior to the OSCE for the PGY3 & 5 residents. No correlation was seen between ITERs and OSCE scores within role, except for the Communicator role (0.64), however, the ITER Overall Scores correlated with the Communicator (0.58), Manager (0.51) and Professional (0.56) OSCE role scores.
<table>
<thead>
<tr>
<th>PD Role Rating</th>
<th>Corresponding OSCE Role Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communicator</td>
<td>0.79</td>
</tr>
<tr>
<td>Collaborator</td>
<td>0.65</td>
</tr>
<tr>
<td>Manager</td>
<td>0.66</td>
</tr>
<tr>
<td>Health Advocate</td>
<td>0.74</td>
</tr>
<tr>
<td>Scholar</td>
<td>0.70</td>
</tr>
<tr>
<td>Professional</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table 10. Correlation between the Program Directors (PD) Ratings of resident ability in each of the Intrinsic Roles, and the corresponding OSCE Role Score.

There was a 64% (16/25) response rate to the resident survey. Overall, 87.6% agreed or strongly agreed that the scenarios reflected encounters that an orthopedic surgeon would have to deal with in their general practice, while 81.3% agreed or strongly agreed that participating in the OSCE would help prepare them for their final Royal College examination. However, only 56.3% agreed or strongly agreed that the OSCE was an effective way to assess their understanding of each of the CanMEDS roles.

**DISCUSSION**

This orthopedic OSCE, designed to test the six intrinsic CanMEDS roles, has shown excellent overall reliability, as well as excellent reliability for the Roles of Communicator, Collaborator, Professional and Manager. Furthermore, using role specific global ratings we have demonstrated an ability to distinguish between orthopedic residents with different years of training. This is the first time formal
assessment of intrinsic roles has been studied in the field of post-graduate orthopedic surgical training.

With regards to total test scores, each of the stations were able to demonstrate a statistically significant difference by year of training, with the exception of Station 3. Station 3 was a case based scenario using a standardized patient – a grandmother who has brought in a child thought to have sustained a non accidental injury. Residents were asked to take a focused history regarding the home situation, and explain to the grandmother the need to alert the appropriate authorities and admit the child. Despite the fact that the PGY effect on station 3 scores was not statistically significant (p = 0.068), the scores demonstrate the same general trend as all other stations, i.e., PGY5 (mean score = 90.5%) > PGY3 (mean score = 74.3%) > PGY1 (mean score = 69.5%). It may be that this station did not achieve significance because the PGY0s and PGY3s performed well, suggesting that these competencies may have been covered in undergraduate medical programs.

Careful blueprinting was used in this CanMEDS OSCE to avoid redundancy – roles were spread between stations, and those stations that did assess the same roles focused on different competencies within that role, as outlined in the CanMEDS 2005 Framework (Frank 2005). For example, in this OSCE, two stations (Needlestick and Trauma List) both examined the roles of Communicator and Manager, with the Needlestick station additionally examining resident understanding of the Professional role (bioethical principles and informed consent). However, the Trauma List station focused on the competencies of priority setting and time management within the Manager role, while the Needlestick station sought to examine the competency of managing practice and career effectively.

An OSCE has been previously been shown to be a valid and reliable tool for the assessment of Medical Expert, with some evidence for its use in assessing the role of Communicator (Srinivasan 1999, Keely, Myers et al. 2002, Yudkowsky, Alseidi et al. 2004). Improved communication skills have previously been linked to both advanced year of training (Warf, Donnelly et al. 1999), and to increased clinical competence (Colliver, Swartz et al. 1999). An OSCE has also been adapted to assess
competencies within the roles of Professional (Hilliard 2000, Altshuler and Kachur 2001, Singer, Pellegrino et al. 2001) and Scholar (Schol 2001, Fliegel, Frohna et al. 2002), with varying amounts of success. For example, Singer et al. in an OSCE designed to assess clinical ethics, found a low reliability with only four stations; it was felt that increasing the number of stations would be required to obtain acceptable reliability (Singer, Pellegrino et al. 2001).

Jefferies et al. recently demonstrated that an OSCE may be a valid and reliable method of simultaneously assessing multiple competencies in neonatal-perinatal medicine (Jefferies, Simmons et al. 2007). Subspecialty trainees were assessed using a combination of binary checklists, 5-point CanMEDS ratings, as well as SP’s and SHP’s assessment of interpersonal and communication skills. Inter-station reliability was acceptable to excellent for six of the seven roles, with the exception of Scholar. Only the teaching component of Scholar was assessed – the authors recommended creating a single station to assess the competencies inherent to the Scholar role, including the ability to understand and evaluate research. We applied this technique with success – our Scholar station, designed to assess application of evidence based medicine, was able to distinguish between residents with different levels of training.

Jefferies et al. also looked at the use of the structured oral examination in assessing the seven CanMEDS roles including Medical Expert – interstation reliability was acceptable for the roles of Medical Expert, Scholar and Professional (between 0.6 and 0.8), but not for Communicator, Collaborator and Health Advocate (between 0.4 and 0.6) or for Manager (0.19) (Jefferies, Simmons et al. 2011). In comparison to their previous OSCE study, inter-station reliability was lower for all roles except for Scholar. However, costs were reduced significantly by not using standardized patients. While we felt that SP’s were an important component of our OSCE, the costs (in the region of $3000) were not insignificant, equating to a cost of $250 per resident – it may be possible to substitute orthopedic fellows or staff in place of SP’s in future iterations. However, given the importance of establishing competence in these areas by both the Royal College of Surgeons and Physicians Canada, and the American Council for
Graduate Medical Education (ACGME), this could be seen as a reasonable cost for training programs to bear on an annual basis.

The Royal College of Physicians and Surgeons of Canada (RCPSC) has a published handbook detailing assessment methods for the CanMEDS Competencies (Bandiera 2006). This document states that oral examinations and OSCEs are not well suited to evaluate the roles of Manager and Scholar. Other documents attest to the perceived difficulty with assessing the intrinsic roles, especially Health Advocate (Verma, Flynn et al. 2005, Chou, Cole et al. 2008, Murphy, Bruce et al. 2008). However, the reliability of Manager in our study was high enough to be used in a high stakes examination. While we cannot attest to the reliability of Health Advocate and Scholar due to insufficient items, ANOVA testing demonstrated a significant ability to distinguish between residents of different PG years of training in both of these roles. We believe that in fact the OSCE is a very appropriate means of assessment, as clinical scenarios can be used that mimic real life encounters.

An advantage of this type of OSCE is that both teaching (formative evaluation) and assessment (summative evaluation) can be incorporated. As noted by Zuckerman et al., assessment motivates residents to learn important skills, and is therefore a form of learning in itself (Zuckerman, Holder et al. 2012). We believe that by exposing very junior residents (PGY0) to scenarios they will likely soon encounter (complaints of delayed surgery, difficult interactions with operating room staff), learning opportunities can be created in an environment suitable for feedback and coaching (Duffy, Gordon et al. 2004). Furthermore, by retesting mid rank residents (PGY3), an assessment of their skills in each of the CanMEDS roles can be re-evaluated, and appropriate feedback provided. At our institution, a bank of multiple CanMEDS scenarios has been created; we believe that all residents will benefit from exposure to a CanMEDs OSCE twice in their training, once as a junior and once as a senior.

We are not aware of any OSCE designed to test only the intrinsic CanMEDS roles. While it is difficult to remove Medical Expert from such an examination, every effort was made to minimize scenarios dependent on orthopedic knowledge. For example, in the station focusing on the role of Manager, residents were asked to manage a
overbooked trauma list; some degree of orthopedic knowledge was required to know the urgency of each case, but residents were graded on their reasoning, and on their ability to handle a phone call from a disgruntled relative. In the Scholar station, residents were expected to know levels of evidence, and how to perform database searches; in the Needlestick case (Professional) residents were expected to know the immediate and delayed management of such an occurrence, as well as the ethical principles involved regarding patient consent, and notification of the appropriate monitoring bodies. For this reason, we do not believe that there were any major qualitative differences regarding the degree of core knowledge assessed in each station.

We were interested in obtaining concurrent validity; for example how could the station creator be certain that communication stations were truly assessing the communicator role. All case scenarios were based upon real-life clinical situations, and adherent to role descriptions provided by the Royal College of Physicians and Surgeons Canada (Frank 2005, Jefferies, Simmons et al. 2007). Interestingly, no correlation was seen between the ITER role scores and the OSCE role scores, but a good correlation was seen with Program Director ratings of the roles. This suggests that ITERs are not a particularly effective form of assessment for the intrinsic roles.

Limitations included our inability to comment on the reliability of the roles of Scholar and Health Advocate, as only a single global rating was used for each of these roles – this will be remedied in the future. However, each of these roles was useful in distinguishing between different years of training. Objectively may have been increased by the use of SP’s or SHP’s to provide global ratings of the residents, a method which has been used to good effect in the medical education literature, with evidence of good correlation between ratings completed by SP’s and faculty physicians (Cooper and Mira 1998, Colliver, Swartz et al. 1999, Donnelly, Sloan et al. 2000). Importantly, the examiners will have known some residents and their PGY of training, raising the potential for bias. Examiners were asked to disregard the PGY level of the resident – however it may be that the use of SP’s ratings will help to offset this risk. In this OSCE, neither station nor role weighting was used, as it was felt that
each of the CanMEDS roles was equally important. Finally, it is not possible to know how this CanMEDS OSCE compares to a more traditional OSCE with incorporated assessment of CanMEDS roles within those stations – however, we have demonstrated a high degree of reliability or internal consistency, one measure that an exam is performing well. It may be that the high degree of reliability seen in this CanMEDS OSCE may be a result of its narrow focus.

CONCLUSION

An OSCE designed to assess the intrinsic CanMEDS roles proved to be sufficiently valid and reliable to be used regularly in an orthopedic residency program.
This chapter presents the results of an OSATS, using simulation and dry models to assess the performance of technical procedures by residents after a sports medicine rotation. Once again, a significant difference was seen in performance between junior and senior residents - less than 50% of junior residents were able to achieve a minimal level of competency on any of the technical procedures, despite having been trained in these procedures in the operating room and on the models. However, in contrast to the clinical skills OSCE, this study demonstrates that many senior residents also have difficulty in achieving a minimal level of competence in some of the technical procedures.
ABSTRACT

Background
As post-graduate medical training shifts to a competency-based model, methods of assessing performance of technical procedures become necessary. We hypothesized that an Objective Structured Assessment of Technical skill (OSATS), using dry models, would be a valid method of assessing residents' ability to perform sports medicine procedures. We also hypothesized that, after training in a Competency-based Model of Education (CBME) program, junior and senior residents would be able to perform technical procedures to a similar level.

Methods
Each resident was provided a list of 10 procedures in which they were expected to demonstrate competence. At the end of the rotation, each resident undertook an OSATS comprised of six stations using dry models – faculty used the Arthroscopic Surgical Skill Evaluation Tool (ASSET)), task-specific checklists, as well as an overall five-point global rating scale (GRS) to score each resident. Each procedure was videotaped for blinded review.

Results
Over 18 months, 27 residents (19 junior (PGY1-3), 8 senior (PGY4&5)) sat the OSATS after their rotation, as well as 14 sports medicine staff and fellows. The reliability of the OSATS was high (0.9), as was the correlation with blinded observer review of videotapes (>0.8). A significant difference by year in training was seen for the overall GRS, the total ASSET score, and the total checklist score, as well as for each technical procedure (p<0.001). Further analysis demonstrated a significant difference in the total ASSET score between junior and senior residents, senior residents and fellows, as well as between fellows and faculty (p<0.05).

Conclusions
The results of this study demonstrate that an OSATS using dry models shows
evidence of validity when used to assess performance of technical procedures after a sports medicine rotation. However, junior residents were not able to perform as well as senior residents, suggesting that overall surgical experience is as important as intensive teaching.

INTRODUCTION

The acquisition of competent surgical skill is a defining characteristic of surgery, but is not measured systematically in residency (Williams, Verhulst et al. 2012). As competency-based medical education (CBME) continues to become more prevalent across North America, objective assessment tools are required in order to demonstrate competency. By using regular assessment to ensure a minimal level competence is achieved by the conclusion of a resident’s rotation (e.g., sports medicine), before progression to the next rotation, orthopedic education will move from a time-based program to one based upon observable and measurable outcomes (Pugh, Touchie et al. 2014).

Previous work has validated methods for assessing medical knowledge (Dwyer, Theodoropoulos et al. 2013), and communication skills (Dwyer, Glover Takahashi et al. 2014) in orthopedic residents. However, valid, reliable and objective methods of assessing competence in the performance of technical procedures have yet to be fully defined in orthopedics. While the gold standard will likely always be in the operating room, difficulty standardizing operations (Reznick, Regehr et al. 1997), issues of patient safety, clinical outcomes, as well as time restraints (Farnworth, Lemay et al. 2001), mean that residents often cannot perform complex procedures to completion in that environment. As such, we believe that simulation-based assessment may offer part of the solution to this complex problem.

At our university, residents undertake two three-month orthopedic sports medicine rotations during their training, once as a junior (postgraduate year (PGY) 1 – 3) and once as a senior (PGY 4 – 5). During these rotations, their education is comprised of
clinical and operating room exposure, regular didactic lectures based upon a set curriculum, and the teaching of technical procedures in a laboratory utilizing anatomical knee and shoulder dry models. The skills laboratory is also available for use on an unlimited, independent basis throughout the rotation. At the end of the rotation, medical knowledge is assessed with an Objective Structured Clinical Examination (OSCE) (Dwyer, Theodoropoulos et al. 2013), and technical skill with an Objective Structured Assessment of Technical Skills (OSATS) using dry models.

The purpose of this study was to evaluate the validity of an OSATS using simulation to assess performance of technical procedures after a sports medicine rotation. We hypothesized that a simulated OSATS using dry models would show evidence of validity, suitable for use as an in-training assessment of competence after a sports medicine rotation. We also hypothesized that, after training within a CBME model, junior and senior residents would be able to perform technical procedures at a similar level.

MATERIALS AND METHODS

In a prospective study, all residents who undertook the sports medicine rotation were included in this study. At the beginning of the rotation, each resident was provided with a list of 10 technical procedures that they were expected to demonstrate competency in by the end of the rotation (Table 11). The surgical procedures were selected by a focus group of academic orthopedic sports surgeons, and corresponded with the sports medicine curriculum. At the end of each rotation, six of these technical procedures were selected for assessment in an OSATS, which was comprised of six stations, each ten minutes long. Station selection was performed in order to allow reasonable representation of all stations across all groups of participants.
<table>
<thead>
<tr>
<th>Task</th>
<th>Component of task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perform partial meniscectomy</td>
<td>Radial tear created in the posterior horn of medial meniscus</td>
</tr>
<tr>
<td></td>
<td>Resident directed to perform partial medical meniscectomy</td>
</tr>
<tr>
<td>Whipstitch hamstring graft</td>
<td>Two equal lengths of shoestring provided</td>
</tr>
<tr>
<td></td>
<td>Resident directed to whipstitch each end and pretension graft</td>
</tr>
<tr>
<td></td>
<td>on Smith and Nephew Graftmaster II (Andover, USA)</td>
</tr>
<tr>
<td>Drill anteromedial ACL femoral tunnel</td>
<td>Resident directed to drill femoral tunnel using anteromedial portal, with 8mm tunnel and with 15 mm Endobutton (Smith and Nephew, Andover, USA) femoral fixation</td>
</tr>
<tr>
<td>Drill transtibial ACL femoral tunnel</td>
<td>Resident directed to drill femoral tunnel using transtibial tunnel</td>
</tr>
<tr>
<td></td>
<td>(tibial predrilled with size 10 mm tunnel)</td>
</tr>
<tr>
<td>Drill ACL Tibial Tunnel</td>
<td>Resident directed to drill tibial tunnel</td>
</tr>
<tr>
<td>Insert glenoid anchors</td>
<td>Resident directed to insert 2.3 mm glenoid anchor (Bioraptor, Smith and Nephew, Andover, USA) at 5 o’clock position</td>
</tr>
<tr>
<td>Pass labral sutures</td>
<td>Resident directed to pass labral sutures using Accu-Pass (Smith and Nephew, Andover, USA) suture passing device</td>
</tr>
<tr>
<td>Insert rotator cuff anchors</td>
<td>Resident directed to insert 5 mm metal Twinfix anchor (Smith and Nephew, Andover, USA) into footprint of humeral head</td>
</tr>
<tr>
<td>Pass rotator cuff suture</td>
<td>Resident directed to pass two sutures through tendon. Both Scorpion and Scorpion Fast Pass available for use (Arthrex, Naples, USA).</td>
</tr>
<tr>
<td>Tie sliding, locking arthroscopic knot</td>
<td>Faculty insert rotator cuff anchor, and pass sutures prior to performance of knot tying (Shoulder Arthroscopy Model)</td>
</tr>
</tbody>
</table>

Table 11. List of technical procedures that residents were expected to demonstrate a minimal level of competence at after a sports rotation. ACL – anterior cruciate ligament.

The majority of shoulder procedures (insertion glenoid anchors, labral suture passage, insertion rotator cuff anchors, rotator cuff suture passage) were performed on the Arthroscopic SOFT Shoulder Model (model DS-012, Arthrex, Naples, USA) bench top simulator using replacement joint capsule assembly with cuff tear (DS-012-3SST) (Figure 6). Faculty inserted portals into the subacromial space or the glenohumeral joint prior to the assessment. Knot tying was performed on a Shoulder Arthroscopy Model (S.A.M., www.arthrodemo.com) – faculty placed a rotator cuff anchor, and passed both suture limbs through the tissue prior to performance of a sliding, locking arthroscopic knot.
Components of ACL reconstruction and partial meniscectomy were performed using an ACL Sawbones model (Sawbones, Washington, United States), utilizing an encapsulated knee insert (model 1414-1) inside a soft tissue with skin (model 1413-1) (Figure 7). Again, all portals were made prior to the assessment. Standard 30° arthroscopic cameras with high definition video systems were used for all procedures, in association with disposable implants and standard equipment (Smith & Nephew, Andover, USA) (Arthrex, Naples, Fl, USA).
Residents were evaluated using a combination of task-specific checklists, a previously validated global rating scale (GRS) (the Arthroscopic Surgical Skill Evaluation Tool (ASSET)) (Koehler, Amsdell et al. 2013), and a final five-point GRS corresponding to the Drefus model of skill acquisition (novice, advanced beginner, competent, proficient, expert) (Batalden, Leach et al. 2002, Carraccio, Benson et al. 2008). Task-specific checklists were created for each technical procedure using established arthroscopy texts, and by a panel of experienced sports medicine staff surgeons – consensus was achieved using a modified Delphi procedure conducted by way of multiple surveys (Koehler, Amsdell et al. 2013). The ASSET is made up of eight domains (safety, field of view, camera dexterity, instrument dexterity, bimanual dexterity, flow of procedure, quality of procedure, and autonomy) based upon the Dreyfus model of skill acquisition (Koehler, Amsdell et al. 2013). The ASSET was designed to be generalizable to multiple procedures and settings (operating room and simulation), and has been found to be reliable and valid in the setting of diagnostic knee arthroscopy in cadaveric specimens (Koehler, Amsdell et al. 2013).

A pilot OSATS was run using two junior residents, in order to establish the practicality of each station (sufficient time to complete each task) and optimize the checklist for
each station. All residents at the end of the sports rotation undertook the OSATS; all staff surgeons with fellowship training in sports medicine at our institution, and all sports medicine fellows also sat the OSATS. A single examiner was present at each station – examiners were orthopedic sports medicine surgeons, who sat the OSATS prior to becoming examiners. All examiners were familiar with the checklists and the GRS, and were available to assist as directed by or requested by residents. No guidance with regards to the procedure was provided at any time. Videotaping of all procedures (both hand movements and arthroscopic component) was performed, and reviewed by a blinded, independent sports medicine fellow.

**Statistical Analysis**

STATA version 13 (College Station, Texas) was used to perform all analyses. The total score of the ASSET was calculated as the sum of the eight domains, with a maximum score of 38. The internal consistency / reliability of the total ASSET score, total checklist score, and the overall GRS was assessed using Cronbach’s Alpha. Individual station reliability was calculated using Cronbach’s Alpha if Item Deleted, which is calculated by removing each station before recalculating the overall reliability – if removing any one station increases the Cronbach’s Alpha it implies that that station is performing poorly. Construct validity (the ability to discriminate between different years of training) was assessed using one-way analysis of variance (ANOVA), with year of training (junior resident, senior resident, fellow, staff) as the independent variable. Unpaired t-tests were used to examine for differences between groups. Concurrent validity was sought by examining the correlation between the end of rotation OSCE results for each of the residents, and the OSATS, using Pearson’s correlation. Inter-rater reliability between the examiners and blinded video reviewer was calculated using the intra-class correlation coefficient (ICC) for total ASSET score. Reliability was also calculated using generalizability theory, analyzed using G-String IV software (Bloch & Norman, Hamilton, ON, Canada), using common station combinations, and estimating the variance created by participants, their training levels, stations, as well as items within stations.
A sample size calculation was performed; a minimum of eight residents was required in each group (junior, senior) to detect a difference of 1.0 in the overall GRS, using a mean expected global rating of 3 (competent), standard deviation of 1.0, alpha of 0.05, and a power of 0.8. Costs for equipment and models were also calculated per participant, excluding implants which were donated by industry.

**Ethics**

This study was conducted after approval by the institutional research ethics board.

**Source of funding**

This study was supported by a Core Competency Innovation Grant from the OMeGA Medical Grants Association.

**RESULTS**

Over 18 months, 27 residents (19 junior (PGY1-3), eight senior (PGY4&5)) sat an OSATS after their rotation, as well as seven sports medicine staff and seven fellows, for a total of 41 participants. No resident sat the exam both as a junior and as a senior. The internal consistency / reliability (Cronbach’s Alpha) of the OSATS was excellent (0.95). A good correlation was seen between the overall checklist score and the overall global rating (0.71). For no station did the Cronbach’s Alpha if Item Deleted increase, indicating that all stations were performing well.

Analysis using ANOVA demonstrated a significant difference by groups for the overall GRS, the total ASSET score, and the total checklist score for all stations combined (p<0.001) (Figure 8). A significant difference (ANOVA) was also seen for the ASSET score for each individual station by level of training (Figure 9) (all p<0.001). Analysis using a t-test demonstrated a significant difference in the total ASSET score between junior and senior residents, senior residents and fellows, as well as between fellows and staff (p<0.001). The number of participants deemed competent (overall GRS of competent / proficient / expert) for each technical procedure is seen in Tables 12 & 13.
Figure 8. (A) Box plot for the overall checklist scores calculated as a percentage. (B) Box plot for the total ASSET global rating. (C) Box plot for the overall global rating. There was a significant difference by year of training for each (p<0.001).
Figure 9. Boxplots for the total ASSET score for each of the stations. A significant difference by year of training was seen for all technical skills.
Table 12. Number and percentage of participants deemed competent or better in each knee station using the overall global rating scale (novice, advanced beginner, competent, proficient, expert).

<table>
<thead>
<tr>
<th>Participants</th>
<th>Anteromedial Femoral Tunnel</th>
<th>Transtibial Femoral Tunnel</th>
<th>Tibial Tunnel</th>
<th>Partial Meniscectomy</th>
<th>Whipstitch Hamstring Graft</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junior residents</td>
<td>1/9 (11%)</td>
<td>1/8 (12.5%)</td>
<td>4/14 (29%)</td>
<td>4/9 (44%)</td>
<td>5/12 (42%)</td>
</tr>
<tr>
<td>Senior residents</td>
<td>3/5 (60%)</td>
<td>3/3 (100%)</td>
<td>5/6 (83%)</td>
<td>5/6 (83%)</td>
<td>5/5 (100%)</td>
</tr>
<tr>
<td>Fellows</td>
<td>4/4 (100%)</td>
<td>3/3 (100%)</td>
<td>3/4 (75%)</td>
<td>7/7 (100%)</td>
<td>6/7 (86%)</td>
</tr>
<tr>
<td>Staff</td>
<td>5/5 (100%)</td>
<td>3/3 (100%)</td>
<td>7/7 (100%)</td>
<td>7/7 (100%)</td>
<td>4/4 (100%)</td>
</tr>
</tbody>
</table>

Table 13. Number and percentage of participants deemed competent or better in each shoulder station using the overall global rating scale (novice, advanced beginner, competent, proficient, expert).

<table>
<thead>
<tr>
<th>Participants</th>
<th>Insertion glenoid anchors</th>
<th>Pass labral sutures</th>
<th>Insertion rotator cuff anchors</th>
<th>Pass rotator cuff suture</th>
<th>Tie arthroscopic knot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junior residents</td>
<td>4/11 (36%)</td>
<td>2/11 (18%)</td>
<td>5/13 (38%)</td>
<td>5/13 (38%)</td>
<td>3/16 (19%)</td>
</tr>
<tr>
<td>Senior residents</td>
<td>2/5 (40%)</td>
<td>2/5 (40%)</td>
<td>4/4 (100%)</td>
<td>2/4 (50%)</td>
<td>5/7 (71%)</td>
</tr>
<tr>
<td>Fellows</td>
<td>3/3 (100%)</td>
<td>3/3 (100%)</td>
<td>4/4 (100%)</td>
<td>4/4 (100%)</td>
<td>7/7 (100%)</td>
</tr>
<tr>
<td>Staff</td>
<td>3/3 (100%)</td>
<td>3/3 (100%)</td>
<td>4/4 (100%)</td>
<td>4/4 (100%)</td>
<td>6/6 (100%)</td>
</tr>
</tbody>
</table>

The inter-rater reliability between the examiners ratings, and the blinded video reviewer for the total ASSET score using ICC was 0.9. A good correlation was seen between performance on the end of rotation OSCE (overall GRS summed over the six stations), and each resident’s performance total ASSET score on the OSATS (0.76). A generalizability coefficient was calculated using the knot tying, partial meniscectomy, and whip stitching stations, as well as the stations of glenoid anchor insertion and labral suturing – the relative error coefficient (interstation reliability) was 0.64 and 0.57 respectively. Using these calculations, it was estimated that a six station exam would confer an overall reliability of 0.78. The major source of variance (50%) was training level of the participant.
During this study, 39 Sawbones knees ($69.75 each) with five knee skins ($83.70 each) were used for components of ACL reconstruction and meniscectomy. For components of rotator cuff and labral repair, four Arthroscopic SOFT Shoulder Models ($325) and 25 replacement inserts ($85) were required, with knot tying was performing using a single S.A.M model ($495) with 18 basic rotator cuff felt components ($6) and 18 humeral head replacements ($25). The total cost was $7,616.75, which averaged $185 for the OSATS per participant.

**DISCUSSION**

The most important finding of this study is that an OSATS shows evidence of validity, when used to assess the performance of technical procedures by residents on dry models after a sports medicine rotation. The results also show that junior residents are not able to perform these simulated procedures as well as senior residents, suggesting that a combination of increased exposure to sports medicine, and overall surgical experience is a factor in the acquisition of these technical skills.

A recent survey of orthopedic program directors and residents identified that a significant number believed that surgical skills simulation should become a required part of training (Karam, Pedowitz et al. 2013). Simulation-based training and assessment allows the opportunity for residents to attempt complex surgical tasks, without risk to patients. Simulation also allows residents the opportunity to make independent, intra-operative decisions, taking responsibility for and managing the consequences of their actions – tangled sutures and improperly inserted anchors are one example. Furthermore, simulation provides a staff surgeon with valuable insight into each resident’s knowledge of procedures, and their familiarity with instrumentation prior to performance in the operating room.

Options for simulation in orthopedics include high fidelity models (cadavers), virtual reality, and low fidelity models (sawbones). Cadavers are thought to be the gold standard for simulation training (Reznick, Regehr et al. 1997, Anastakis, Regehr et al.
1999), but require advanced facilities and have significant costs – because of this, we believe that cadavers are not a feasible option for regular assessment of technical procedures within a CBME model of training. Virtual reality (VR) simulation is commercially available for knee and shoulder arthroscopy – studies have demonstrated an ability to distinguish between differing skill levels, as well as an ability to demonstrate improvement in over time (Pedowitz, Esch et al. 2002, Srivastava, Youngblood et al. 2004, McCarthy, Moody et al. 2006, Gomoll, O'Toole et al. 2007, Gomoll, Pappas et al. 2008). However, VR has limited ability to allow the performance of complex procedures such as the insertion of anchors in the humeral head or glenoid, or the drilling of ACL tunnels.

The advantages of dry models are many, including relative ease of preparation, reduced cost, and the opportunity for unsupervised practice (Butler, Olson et al. 2013). The majority of comparative studies have demonstrated that, overall, low-fidelity simulators are similarly effective but less expensive than high-fidelity simulators with regards to the acquisition of surgical skills (Matsumoto, Hamstra et al. 2002, Grober, Hamstra et al. 2004, Chandra, Savoldelli et al. 2008, McDougall, Kolla et al. 2009, Zendejas, Wang et al. 2013). Studies have also shown that the use of low-fidelity models improves the performance of procedures such as shoulder and knee arthroscopy in the operating room (Howells, Gill et al. 2008, Alvand, Auplish et al. 2011, Butler, Olson et al. 2013).

The finding that many residents were unable to perform the sports medicine procedures in the simulated setting warrants attention. This finding is task dependent – while most of the procedures were difficult for the junior residents, some procedures were also difficult for senior residents to complete. This was especially noticeable with the simulated arthroscopic shoulder procedures – one reason may be that these are advanced skills, requiring significant experience and / or fellowship training. It may also highlight deficiencies in our training program - in the operating room staff surgeons may be reluctant to allow residents to perform these procedures, due to the complexity of the individual steps.
The finding that there is a significant difference in performance between junior and senior residents is at odds with the concept of CBME, which in theory demands a minimal level of competence prior to residents proceeding to the next rotation. How to best deal with these results is unknown. One approach would be to increase the frequency of teaching in the simulation laboratory - however these are already being held twice monthly over the three-month rotation. These training sessions are not without cost, both in terms of faculty time, and costs of models and implants – for this reason the number of skills sessions cannot be easily increased. Alternatively, a milestones approach, as recommended by the ACGME, could be implemented (ACGME 2013). In this way, while teaching and training sessions would continue to cover the surgical procedures listed, junior residents would be expected to demonstrate a minimal level of competency at less complex tasks (partial meniscectomy, whipstitch hamstring graft), while it seems that senior residents can be expected to perform aspects of ACL reconstruction and rotator cuff repair.

Importantly, this study does not provide any correlation between the procedures performed in the simulated setting, and operative performance. While there are some studies demonstrating that bench model training improves operative performance in orthopedics (Howells, Gill et al. 2008, Butler, Olson et al. 2013), the evidence that simulation-based assessment correlates with actual surgical performance is more limited (Korndorffer, Kasten et al. 2010), although studies in other surgical specialties have consistently identified good correlations (Datta, Bann et al. 2004, Beard, Jolly et al. 2005, McCluney, Vassiliou et al. 2007, Ghaderi, Vaillancourt et al. 2011). The reasons for this deficiency in the literature is likely because of difficulty standardizing operations, time restraints, and issues of patient safety (Reznick, Regehr et al. 1997, Farnworth, Lemay et al. 2001). While there is no reason to believe that residents who cannot perform these procedures in the simulated setting would be able to in the workplace, this is an important area of future research in CBME.

The OSATS study identified a correlation between performance of technical procedures, and the OSCE scores for each resident. However, the results of other
studies seeking a correlation between medical knowledge and technical skills have been mixed (Bann, Khan et al. 2004) - a recent study looking at simulated surgery on a distal radius fracture failed to identify a correlation between knowledge testing, and biomechanical testing of the surgical construct (Putnam, Kinnucan et al. 2015). In 2009, Van Heest et al. reported on orthopedic residents performing a carpal tunnel release on a cadaver, in association with a knowledge test – while a knowledge score < 70/100 predicted failure on the OSATS, a score of > 70/100 did not ensure that the examinee would pass the test (Van Heest, Putnam et al. 2009). These findings are plausible – while there is threshold of knowledge that is required, above that threshold, a combination of technical abilities and surgical experience has an important role.

The most important limitation of this study is that we are unable to demonstrate a correlation between performance on this OSATS study, and performance in the operating room. Furthermore, while residents were exposed to training on dry models, inevitably each will have had different clinical exposures during their rotations – it is also unknown how much each resident took advantage of the skills lab outside of teaching sessions. The number of sports medicine and other procedures each resident had experience with is unknown – for this reason we cannot say whether these results are a result of sports medicine experience, overall surgical experience, or a combination of both.

CONCLUSION

The results of this study demonstrate that an OSATS using dry models shows evidence of validity when used to assess performance of technical procedures after a sports medicine rotation. However, junior residents were not able to perform as well as senior residents, suggesting that overall surgical experience is as important as intensive teaching.
Chapter 8
Cognitive and Psychomotor Entrustable Professional Activities: Can Simulators Help Assess Competency in Trainees?

Dwyer T, Wadey V, Archibald D, Kraemer W, Shantz JS, Townley J, Ogilvie-Harris D, Petrera M, Ferguson P, Nousiainen M.
Clin Orthop Relat Res. 2015 Sep 22
PMID: 26394640

This chapter presents the results of a study examining assessment of Entrustable Professional Activities (EPAs) using simulation, focusing on pre- and postoperative management of patients (OSCE format), and the performance of technical procedures (OSATS). Mapped to the top 10 EPA’s as designated by faculty, PGY-1 and PGY-4 residents were assessed on their ability to perform three EPAs; managing a patient with hip fracture, ankle fracture, and a patient for a total knee replacement. All of the PGY-1 residents involved in this study were at the end of the first year of training, and each had been deemed competent in the CBME rotations of basic trauma, hip fracture, and basic arthroplasty. The performance of the PGY-1 residents was then compared to PGY-4 residents – the majority of the PGY-4 residents had recently finished complex trauma and arthroplasty rotations. On all of the three technical procedures stations, and most of the patient management stations, junior residents were not able to perform as well as senior residents.
Abstract

Background An entrustable professional activity describes a professional task that postgraduate residents must master during their training. The use of simulation to assess performance of entrustable professional activities requires further investigation.

Questions/purposes (1) Is simulation-based assessment of resident performance of entrustable professional activities reliable? (2) Is there evidence of important differences between Postgraduate Year (PGY)-1 and PGY-4 residents when performing simulated entrustable professional activities?

Methods Three entrustable professional activities were chosen from a list of competencies: management of the patient for total knee arthroplasty (TKA); management of the patient with an intertrochanteric hip fracture; and management of the patient with an ankle fracture. Each entrustable professional activity was 40 minutes long with three components: preoperative management of a patient (history-taking, examination, image interpretation); performance of a technical procedure on a sawbones model; and postoperative management of a patient (postoperative orders, management of complications). Residents were assessed by six faculty members who used checklists based on a modified Delphi technique, an overall global rating scale, as well as a previously validated global rating scale for the technical procedure component of each activity. Nine PGY-1 and nine PGY-4 residents participated in our simulated assessment. We assessed reliability by calculating the internal consistency of the mean global rating for each activity as well as the inter-rater reliability between the faculty assessment and blinded review of videotaped encounters. We sought evidence of a difference in performance between PGY-1 and PGY-4 residents on the overall global rating scale for each station of each entrustable professional activity.

Results The reliability (Cronbach’s α) for the hip fracture activity was 0.88, 0.89 for the ankle fracture activity, and 0.84 for the TKA activity. A strong correlation was seen between blinded observer video review and faculty scores (mean 0.87 [0.07], p < 0.001). For the hip fracture entrustable professional activity, the PGY-4 group had a higher mean global rating scale than the PGY-1 group for preoperative management.
(3.56 [0.5] versus 2.33 [0.5], p < 0.001), postoperative management (3.67 [0.5] versus 2.22 [0.7], p < 0.001), and technical procedures (3.11 [0.3] versus 3.67 [0.5], p = 0.015). For the TKA activity, the PGY-4 group scored higher for postoperative management (3.5 [0.8] versus 2.67 [0.5], p = 0.016) and technical procedures (3.22 [0.9] versus 2.22 [0.9], p = 0.04) than the PGY-1 group, but no difference for preoperative management with the numbers available (PGY-4, 3.44 [0.7] versus PGY-1 2.89 [0.8], p = 0.14). For the ankle fracture activity, the PGY-4 group scored higher for postoperative management (3.22 [0.8] versus 2.33 [0.7], p = 0.18) and technical procedures (3.22 [1.2] versus 2.0 [0.7], p = 0.018) than the PGY-1 groups, but no difference for preoperative management with the numbers available (PGY-4, 3.22 [0.8] versus PGY-1, 2.78 [0.7], p = 0.23).

**Conclusions** The results of our study show that simulated EPAs may be used to determine the ability of a resident to perform professional tasks that are critical components of medical training. In this manner, educators can ensure that competent performance of these skills in the simulated setting occurs before actual practice with patients in the clinical setting.

**Introduction**

An entrustable professional activity is a professional task that postgraduate medical and surgical residents must master during their training; that is, these are tasks and responsibilities that faculty entrust a trainee to perform unsupervised once an adequate level of competence has been achieved (ten Cate 2005). Typically, entrustable professional activities consist of units or tasks that make up a management or evaluation process (such as managing a patient with a hip fracture), which, when put together, form the mass of critical elements that define a profession (ten Cate and Scheele 2007). To date, entrustable professional activities have been created for the medical specialties of pediatrics (Jones, Rosenberg et al. 2011), internal medicine (Chang, Bowen et al. 2013, Lowry, Vansaghi et al. 2013, Caverzagie, Cooney et al. 2015), family medicine (Shaughnessy, Sparks et al. 2013), anesthesiology (Jonker, Hoff et al. 2015), and psychiatry (Boyce, Spratt et al. 2011). As of June 2014, the Accreditation Council for Graduate Medical Education required
reporting on selected milestones and incorporating entrustable professional activities into training programs (Nasca, Philibert et al. 2012); however, to date, little research has been carried out on how to best assess resident performance of an entrustable professional activity (Hauer, Soni et al. 2013).

A trial competency-based medical education program was initiated at the University of Toronto in 2009, and as of July 2013, all first-year residents (postgraduate year [PGY]-1) have been automatically enrolled in the competency-based medical education program as part of their medical education and training (Ferguson, Kraemer et al. 2013). Using a process of consensus, our faculty created a list of the top 10 entrustable professional activities for the program (Table 14) from a previously established list of competencies (Wadey, Dev et al. 2009). The ability of orthopedic residents to perform these top 10 entrustable professional activities unsupervised before graduation from our residency program was thought to be critical.
<table>
<thead>
<tr>
<th>Entrustable Professional Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manage a patient with a displaced distal radius fracture</td>
</tr>
<tr>
<td>Manage a patient with a long bone fracture</td>
</tr>
<tr>
<td>Manage a patient with an ankle fracture</td>
</tr>
<tr>
<td>Manage a patient with an intertrochanteric hip fracture</td>
</tr>
<tr>
<td>Manage a patient requiring a below knee amputation</td>
</tr>
<tr>
<td>Manage a patient with a subcapital hip fracture</td>
</tr>
<tr>
<td>Manage a patient with a meniscal tear in the knee</td>
</tr>
<tr>
<td>Manage a patient requiring total knee arthroplasty</td>
</tr>
<tr>
<td>Manage a patient requiring total hip arthroplasty</td>
</tr>
<tr>
<td>Manage a trauma patient requiring transfer to a Level 1 trauma center</td>
</tr>
</tbody>
</table>

**Table 14.** List of the top ten Entrustable Professional Activities at our institution, as determined by faculty.

Options to assess a resident’s ability to perform an entrustable professional activity independently include practice-based assessment and assessment in the simulated setting. In orthopedics, objective structured clinical examinations have been used to determine the ability of residents to manage clinical problems (Dwyer, Theodoropoulos et al. 2013) and to assess their communication and management skills (Dwyer, Glover Takahashi et al. 2014). The ability to perform technical procedures will always be best assessed in the operating room but it can be difficult to standardize surgical procedures (Reznick, Regehr et al. 1997), manage time restraints (Farnworth, Lemay et al. 2001), and ensure optimal patient safety and clinical outcomes (LeBlanc, Hutchison et al. 2013). For these reasons, simulation is increasingly being used to provide opportunities for residents to perform procedures independently, demonstrate knowledge and skills deficits, and commit errors before actually performing surgery on patients in the operating room (Howell 1982, Cannon, Eckhoff et al. 2006, Alvand,
Auplish et al. 2011, Butler, Olson et al. 2013, LeBlanc, Hutchison et al. 2013, D'Angelo, Cohen et al. 2015). At this time, the best method of assessing a resident's ability to perform an entrustable professional activity independently is unknown.

Many entrustable professional activities are longitudinal and describe care that takes place over time (Chang, Bowen et al. 2013). In the clinical setting, components of many entrustable professional activities would require evaluation at separate times and in different settings, making objective assessment difficult. For example, a resident might assess a patient for TKA preoperatively and perform the technical procedure on a different patient later that week. The use of simulation to assess an entrustable professional activity would help overcome these difficulties, allowing faculty to determine a resident’s ability to perform tasks in the clinical setting with the appropriate level of supervision.

We therefore asked: (1) Is simulation-based assessment of resident performance of entrustable professional activities reliable? and (2) Is there evidence of important differences between senior and junior residents when performing simulated entrustable professional activities?

**Materials and Methods**

For our study, from the list of top 10 entrustable professional activities, three entrustable professional activities were selected: management of the patient for TKA; management of the patient with an intertrochanteric hip fracture; and management of the patient with an ankle fracture. The three entrustable professional activities were selected because each is an important component of the first phase of competency-based training at our institution, typically completed within the first year of residency. Furthermore, each of these activities was listed as a key physician competency in the Orthopedic Milestones Project published by the Accreditation Council for Graduate Medical Education (ACGME 2013). Approval for this study was obtained from the institutional research ethics board.
Each entrustable professional activity was 40 minutes long, divided into three parts, each performed at a different station: preoperative management (10 minutes), performance of technical procedure (20 minutes), and postoperative management (10 minutes). The preoperative and postoperative stations followed a previously described and validated Objective Structure Clinical Examination format (Dwyer, Theodoropoulos et al. 2013), focusing on the skills of patient history-taking, physical examination, image interpretation (images displayed on a computer screen), surgical decision-making, obtaining patient consent, and management of patient risk factors (Table 15). History-taking and physical examination were performed on a standardized patient (a trained actor from our standardized patient program) in the TKA activity. No consent from them is necessary, because they are paid to participate. The postoperative management stations involved care of the patient after surgery and management of complications.
<table>
<thead>
<tr>
<th>Entrustable Professional Activity</th>
<th>Preoperative Management</th>
<th>Technical Procedure</th>
<th>Postoperative Management</th>
</tr>
</thead>
</table>

Table 15. The three Entrustable Professional Activities, with a breakdown of the individual components assessed.
The three technical procedure stations were performed using sawbones models (Figure 10). For TKA, residents were asked to perform the distal femoral cut and AP cuts using standard equipment (an industry representative familiar with the equipment was present to guide the specific use of instrumentation). For the intertrochanteric station, a femoral sawbones without a fracture was placed inside a soft tissue cover on a radiolucent table. Residents were instructed to place a sliding hip screw into the femoral head under an image intensifier. For the ankle fracture station, an ankle sawbones in soft tissue was used; residents exposed the fibular fracture, allowing faculty to create a Weber B fracture. Residents were asked to reduce the fracture with a lag screw and obtain stable fixation with a plate.

Figure 10. Technical procedures performed as a component of each entrustable professional activity are shown. (A) ORIF of oblique fibular fracture. (B) Insertion of dynamic hip screw under image guidance. (C) Performance of distal femoral cut and AP cuts for TKA. ORIF = open reduction and internal fixation.
For each of the pre- and postoperative management stations, a checklist was created using a modified Delphi technique with multiple surveys (Koehler, Amsdell et al. 2013). In this manner, a group of content experts (DO-H, WK, MN, PF, JSS, JT, TD, VW) reviewed initial checklists and were asked to add items or alter the wording as required. After this, consensus was achieved whereby reviewers accepted, rejected, or questioned each item; this process occurred until consensus was achieved. Items receiving over 95% consensus were accepted. The checklists were provided to guide expectations at each station.

Examiners also rated the residents using an overall global rating scale based on the Dreyfus model of skill acquisition (novice, advanced beginner, competent, proficient, expert) (Batalden, Leach et al. 2002, Carraccio, Wolfsthal et al. 2002). Examiners were instructed to deem a resident as competent if the resident performed to the level of a qualified orthopaedic surgeon and able to perform this procedure independently without supervision. During the technical procedure, residents were evaluated using a task-specific checklist (also created using a modified Delphi technique) and a previously validated global rating scale designed for use in objective structured assessment of technical skills (OSATS) (Reznick, Regehr et al. 1997, Regehr, MacRae et al. 1998). Examiners were also asked to provide written comments with regard to the performance of each of the technical procedures.

Study participants included nine of 12 available PGY-1 residents who were at the end of their first year of orthopaedic training. Each of the PGY-1 residents was enrolled in the competency-based medical education program and had been deemed competent in the modules of basic arthroplasty, hip and basic fractures, emergency fractures, and management of medical comorbidities in the surgical patient. As a comparison group, nine of 12 available PGY-4 residents were invited to participate; only one of the PGY-4 was enrolled in the competency-based medical education program. All senior residents had undertaken arthroplasty rotations and had recently undertaken a 6-month trauma rotation.

All participants underwent the three entrustable professional activities assessments on the same day. Staff surgeons and fellows served as the examiners (DO-H, WK, MN,
JSS, JT); the same examiner marked each station to maximize consistency. Two examiners marked all three TKA stations for each resident, two examiners marked all of the ankle fracture stations, one examiner marked the hip fracture technical procedure, and one examiner marked the hip fracture pre- and postoperative management stations. During performance of technical procedures, examiners provided assistance as requested but were instructed not to provide feedback.

Examiners were asked to disregard the year of training of the resident, if known, when performing the assessment. To help answer our first question about reliability, six of nine stations (all three ankle stations, the hip fracture technical procedure station, and the ankle fracture preoperative management and technical procedures) were videotaped and reviewed by a blinded observer (MP), allowing inter-rater reliability to be calculated. To answer our second research about differences between PGY-1 and PGY-4 residents, the mean global rating score on each of the three stations of each entrustable professional activity was compared between the two groups. Correlation was also sought between the checklist and the global rating for each station.

*Statistical Analysis*

All data (checklists, global ratings) were deidentified, entered into an Excel spreadsheet (Microsoft Inc, Redmond, WA, USA), and analyzed with the use of SPSS (Version 21; IBM Corp, Armonk, NY, USA). Reliability was calculated using Cronbach \( \alpha \) for the overall global rating scale of the examination (sum of the global ratings for all nine stations converted to a percentage) and for the overall global rating scale for each entrustable professional activity (sum of the global ratings for the three stations of each entrustable professional activity). Individual station reliability was calculated with the use of the Cronbach \( \alpha \) “if item deleted,” whereby the overall reliability was recalculated after removing each station. If removing any station increased the \( \alpha \), it implied that the station was performing poorly. The correlation between the checklist scores and the global rating scale for each station was assessed with the Pearson product moment correlation. A paired t-test was used for analysis of the difference between the two groups of residents. Inter-rater reliability was calculated for each examiner (MN, JSS, DO-H, WK, JT) and the blinded assessor (MP) using an
intracllass coefficient. The number of participants was set (nine in each group); therefore, a power analysis was performed using a t-test with an α value of 0.05, an effect size of 0.5 on the 5-point global rating scale, and a sample size of 18; the power was 0.26.

**Results**

*Reliability of Simulation-based Assessment of Entrustable Professional Activities*

Using performance on the three stations of each entrustable professional activity, internal consistency was 0.84 for the TKA activity, 0.88 for the hip fracture activity, and 0.89 for the ankle fracture activity. The Cronbach α “if item deleted” decreased for every station, demonstrating that each station was performing well (Table 16). A good to high correlation was seen for all nine stations between the checklist scores and the overall global rating scale, suggesting that examiners were using the checklists appropriately (0.72; range, 0.65–0.8; p = 0.01). All videotaped entrustable professional activities showed strong inter-rater agreement with a mean intracllass correlation coefficient of 0.87 (0.07; p < 0.001). Pre- and postoperative management stations of the hip fracture activity as well as the postoperative management station of the TKA were not recorded as a result of feasibility issues.
<table>
<thead>
<tr>
<th>Station</th>
<th>Cronbach’s Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hip fracture preoperative management</td>
<td>0.78</td>
</tr>
<tr>
<td>Hip fracture technical procedure</td>
<td>0.79</td>
</tr>
<tr>
<td>Hip fracture postoperative management</td>
<td>0.75</td>
</tr>
<tr>
<td>Ankle fracture preoperative management</td>
<td>0.8</td>
</tr>
<tr>
<td>Ankle fracture technical procedure</td>
<td>0.78</td>
</tr>
<tr>
<td>Ankle fracture postoperative management</td>
<td>0.79</td>
</tr>
<tr>
<td>TKR preoperative management</td>
<td>0.81</td>
</tr>
<tr>
<td>TKR technical procedure</td>
<td>0.80</td>
</tr>
<tr>
<td>TKR postoperative management</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 16. Cronbach’s Alpha if Item deleted for each station. For no station was this number greater than the overall reliability, indicating that each station was performing well.

**Differences Between Senior and Junior Residents in Simulated Entrustable Professional Activities**

For the hip fracture EPA, the PGY-4 group had a higher mean global rating scale than the PGY-1 group for preoperative management (3.56 [0.5] versus 2.33 [0.5], \( p < 0.001 \)), postoperative management (3.67 [0.5] versus 2.22 [0.7], \( p < 0.001 \)), and technical procedures (3.11 [0.3] versus 3.67 [0.5], \( p = 0.015 \); Table 17). For the TKA activity, the PGY-4 group scored higher for postoperative management (3.5 [0.8] versus 2.67 [0.5], \( p = 0.016 \)) and technical procedures (3.22 [0.9] versus 2.22 [0.9], \( p = 0.04 \)) than the PGY-1 group, but there was no difference for preoperative management with the numbers available (PGY-4, 3.44 [0.7] versus PGY-1 2.89 [0.8], \( p = 0.14 \)). For the ankle fracture activity, the PGY-4 group scored higher for postoperative management (3.22 [0.8] versus 2.33 [0.7], \( p = 0.18 \)) and technical procedures (3.22 [1.2] versus 2.0 [0.7], \( p = 0.018 \)) than the PGY-1 groups, but there
was no difference for preoperative management with the numbers available (PGY-4, 3.22 [0.8] versus PGY-1, 2.78 [0.7], p = 0.23). In general, the majority of PGY-4 residents were able to achieve a level of competency or better in each of the stations; a higher number of PGY-1 residents were not able to, especially in the technical procedure stations (Table 18).

<table>
<thead>
<tr>
<th>Station</th>
<th>PGY 1</th>
<th>PGY4</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hip Fracture Preoperative Management</td>
<td>2.33 (0.5)</td>
<td>3.56 (0.53)</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Hip Fracture Technical Procedure</td>
<td>3.11 (0.33)</td>
<td>3.67 (0.5)</td>
<td>p=0.015</td>
</tr>
<tr>
<td>Hip Fracture Postoperative Management</td>
<td>2.22 (0.67)</td>
<td>3.67 (0.5)</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>TKA Preoperative Management</td>
<td>2.89 (0.78)</td>
<td>3.44 (0.73)</td>
<td>n.s.</td>
</tr>
<tr>
<td>TKA Technical Procedure</td>
<td>2.22 (0.97)</td>
<td>3.22 (0.97)</td>
<td>p=0.04</td>
</tr>
<tr>
<td>TKA Postoperative Management</td>
<td>2.67 (0.50)</td>
<td>3.50 (0.76)</td>
<td>p=0.016</td>
</tr>
<tr>
<td>Ankle Fracture Preoperative Management</td>
<td>2.78 (0.67)</td>
<td>3.22 (0.83)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Ankle Fracture Technical Procedure</td>
<td>2.00 (0.71)</td>
<td>3.22 (1.2)</td>
<td>P=0.018</td>
</tr>
<tr>
<td>Ankle Fracture Postoperative Management</td>
<td>2.33 (0.71)</td>
<td>3.22 (0.83)</td>
<td>P=0.018</td>
</tr>
</tbody>
</table>

Table 17. Mean global rating scale (Standard Deviation) for each of the EPA station components. n.s. – non significant.
<table>
<thead>
<tr>
<th>Station</th>
<th>Number of PGY-1 residents deemed competent</th>
<th>Number of PGY-4 residents deemed competent</th>
<th>Example comments from examiners for residents who where deemed not to be competent at the technical procedure stations</th>
</tr>
</thead>
</table>
| Hip Fracture Preoperative Management | 6/9                                       | 9/9                                       | - No use of guide, free hand, guide wire and screw too posterior  
- Screw unacceptably high in head, reamed to 90, but put 95 screw in  
- No use of guide, free hand, placed wire at 120 degrees  
- Guide wire and screw far too high in head  
- No idea how to assemble triple reamer  
- Generally poor |
| Hip Fracture Technical Procedure | 3/9                                       | 7/9                                       | - No use of guide, free hand, placed wire at 120 degrees  
- Guide wire and screw far too high in head  
- No idea how to assemble triple reamer  
- Generally poor |
| Hip Fracture Postoperative Management | 1/9                                       | 9/9                                       | - No use of guide, free hand, placed wire at 120 degrees  
- Guide wire and screw far too high in head  
- No idea how to assemble triple reamer  
- Generally poor |
| TKA Preoperative Management | 6/9                                       | 8/9                                       | - Used saw with one hand, no retractors, No use of landmarks to set rotation  
- AP cut first, no use of IM rod  
- Constantly flexing blade, unable to stop cutting block popping off  
- Unable to even start TKR, no idea how to do procedure  
- No use of IM guide for distal cut, AP cut first |
| TKA Technical Procedure | 3/9                                       | 8/9                                       | - Used saw with one hand, no retractors, No use of landmarks to set rotation  
- AP cut first, no use of IM rod  
- Constantly flexing blade, unable to stop cutting block popping off  
- Unable to even start TKR, no idea how to do procedure  
- No use of IM guide for distal cut, AP cut first |
| TKA Postoperative Management | 6/9                                       | 8/9                                       | - Used saw with one hand, no retractors, No use of landmarks to set rotation  
- AP cut first, no use of IM rod  
- Constantly flexing blade, unable to stop cutting block popping off  
- Unable to even start TKR, no idea how to do procedure  
- No use of IM guide for distal cut, AP cut first |
| Ankle Fracture Preoperative Management | 6/9                                       | 7/9                                       | - Lag screw did not achieve compression or reduction  
- Poor initial and second lag screw, said malreduced fracture was ok  
- Distal screws far too long and into talus  
- No idea how to perform a lag screw or apply a plate  
- Difficulty with lag screw, caused a fracture, unable to stabilize  
- Fracture unreduced and continued to plate |
| Ankle Fracture Technical Procedure | 2/9                                       | 7/9                                       | - Lag screw did not achieve compression or reduction  
- Poor initial and second lag screw, said malreduced fracture was ok  
- Distal screws far too long and into talus  
- No idea how to perform a lag screw or apply a plate  
- Difficulty with lag screw, caused a fracture, unable to stabilize  
- Fracture unreduced and continued to plate |
| Ankle Fracture Postoperative Management | 6/9                                       | 7/9                                       | - Lag screw did not achieve compression or reduction  
- Poor initial and second lag screw, said malreduced fracture was ok  
- Distal screws far too long and into talus  
- No idea how to perform a lag screw or apply a plate  
- Difficulty with lag screw, caused a fracture, unable to stabilize  
- Fracture unreduced and continued to plate |

**Table 18.** Number of residents deemed competent for each station of each entrustable professional activities.
Discussion

As postgraduate medical education slowly moves toward competency-based medical education, the need to provide objective assessments of competence will continue to be an issue. Creating entrustable professional activities allows faculty to identify and select the most important, representative, and critical tasks that should be mastered (Mulder, Ten Cate et al. 2010). Scheele et al. (Scheele, Teunissen et al. 2008) recommended focusing on those tasks critically important in daily practice or that address high-risk or error-prone activities; certainly the list of competencies listed by the Accreditation Council for Graduate Medical Education include care of patients with ankle fractures, knee osteoarthritis, and hip fractures (ACGME 2013). In such settings, entrustable professional activities may then be used to define five levels of responsibility: observe the activity, act under direct supervision, act under indirect supervision (available within minutes) on call, act unsupervised, and ability to supervise others (ten Cate 2005, ten Cate and Scheele 2007)[36-38]. The ability to achieve level 4 (acting independently) in predetermined entrustable professional activities is a critical component of competency-based medical education (ten Cate 2014). The results of our study demonstrated that simulated activities may be used to determine which residents can perform tasks competently in the simulated setting, allowing these procedures to be performed by residents in the clinical setting under supervision. Most importantly, simulated activities allow for identification of those residents who require further training or remediation to achieve a minimal level of competency. In this way, the simulation of entrustable professional activities can be used effectively to supplement workplace-based assessment of residents.

Our study had a number of limitations. Although we believe that use of simulated patient encounters and simulated technical procedures is valuable, these assessments should be complemented by practice-based assessment. Second, although a simulated assessment of entrustable professional activities was shown to be reliable, validity evidence was limited to the finding that senior residents were able to perform the activities to a higher level than junior residents. Further research is required to demonstrate that the simulated performance of an entrustable professional
activity correlates with actual performance in the clinical setting. Third, only components of each technical procedure were performed as a result of time limitations rather than the entire procedure. For example, it is possible that a resident who was able to perform the femoral cuts of a TKA competently may have had difficulty with the tibial resection. Another important limitation in this study was the potential for rater error or bias, because some of the examiners would have been aware of the year of training of residents. To examine for bias, videotaping was performed on six of the nine stations, and a strong correlation was seen between the examiner ratings and the assessments of the blinded reviewer on these stations. However, not every station was videotaped, so it is not possible to exclude the effect of bias (whereby examiners might be overly stringent on junior residents and overly lenient on senior residents or vice versa) on the findings of differences on the postoperative management of the hip and ankle activities. The study was also rather severely underpowered; for this reason, although we were not able to identify differences between senior and junior residents in the preoperative management of TKA and ankle fracture activity, we cannot exclude the possibility that there was a difference. Finally, simulation was carried out using sawbones models rather than cadavers. The advantages of dry models are many, including relative ease of preparation and reduced cost compared with use of cadavers (Butler, Olson et al. 2013); the majority of comparative studies has also demonstrated that, overall, low-fidelity simulators are similarly effective but less expensive than high-fidelity simulators with regard to the acquisition of surgical skills (Matsumoto, Hamstra et al. 2002, Grober, Hamstra et al. 2004, Chandra, Savoldelli et al. 2008, McDougall, Kolla et al. 2009).

The results of our study demonstrated a high reliability for the entrustable professional activity examination overall for each individual component of the examination and for each station as well as strong inter-rater agreement. However, we have minimal evidence of concurrent validity in this setting or correlation with clinical performance. Interestingly, each of the junior residents had previously been deemed competent in basic arthroplasty and trauma; however, the majority of residents were not assessed specifically on technical procedures such as insertion of a dynamic hip screw or fixation of an ankle fracture in the operating room, and those performing TKA always
did so in the presence of staff providing assistance and feedback, necessary to ensure patient safety and maximize clinical outcomes.

Few studies have evaluated the assessment of entrustable professional activities in postgraduate residency training, either simulated or in the clinical setting. Hauer et al. (Hauer, Soni et al. 2013) conducted two pilot entrustable professional activity-based assessments (inpatient discharge and family meeting) in the clinical setting, testing them on PGY-1 residents in internal medicine. Both the residents and faculty felt the assessments improved skills and facilitated useful feedback. Alyward et al. (Aylward, Nixon et al. 2014) developed a patient handoff entrustable professional activity for interns in internal medicine and pediatrics, identified as a critical skill for residents. Under direct observation, the interns were judged on using the five levels of entrustment as described by ten Cate and Scheele (ten Cate and Scheele 2007) with the majority of residents judged as being able to perform under direct or indirect supervision.

The results of our study showed consistent evidence that senior residents were able to perform most components of each entrustable professional activity at a higher level than junior residents. Although this might be expected, within the competency-based medical education format all junior residents had been previously been deemed competent at each of these activities. Whether there is an issue with skill retention or with the assessment methods used after their previous rotations is unknown. However, we were able to demonstrate deficiencies in both junior and senior residents in various aspects of each simulated activity. Certainly, these findings are consistent with the Orthopaedic Milestones Project, which lists milestone levels that residents will attain as they progress through training (ACGME 2013). For example, in hip fracture patient care, junior residents are expected to move from an ability to take a focused history and perform a focused examination (level 2), to being able to make a comprehensive assessment of fracture patterns and capable of performing surgical repair (level 3) and to being capable of treating postoperative complications such as infection (level 4). Using these competencies as a curriculum guide, the identification of any technical and nontechnical deficits in an objective setting allow for remediation
and reassessment to the advantage of the resident, the program, and future medical practice. Practice-based assessment will always be a critical component, but it is likely desirable that a resident demonstrate an ability to perform a task at a competent level, without supervision, in a simulated environment before working in the operating room or in a clinical setting.

Crucial in medical training, supervision must gradually decrease to build self-confidence and trustworthiness (Hirsh, Holmboe et al. 2014). For supervisors to make valid entrustment decisions, however, sufficient acquaintance of preceptors with trainees, a concept known as “time to trust,” is critical (Hirsh, Holmboe et al. 2014). Kennedy et al. identified that faculty grant residents independence based on the resident’s knowledge and skill as well as their insight into limitations (Kennedy, Regehr et al. 2008). Clearly, long rotations are often required to build sufficient relationships to determine a trainee’s strengths and limitations (Bernabeo, Holtman et al. 2011); brief and fragmented faculty–resident contact is often not an ideal way to draw valid, reliable conclusions (Jones, Rosenberg et al. 2011); however, such close contact between a single resident and faculty can be limited, making a simulated entrustable professional activity valuable in determining a resident’s abilities.

The results of our study show that simulated entrustable professional activities may be used to determine the ability of a resident to perform professional tasks that are critical components of medical training. In this manner, educators can ensure competent performance of these skills in the simulated setting, before actual practice with patients in the clinical setting. Future research needs to demonstrate a correlation between competent performance in the simulated setting and performance in the workplace.
Chapter 9
GENERAL DISCUSSION

9.1 SUMMARY OF FINDINGS

There has been an increasing push for residency training programs to transition from traditional time-based programs, to outcomes-based postgraduate medical education. Nowhere is this change more evident than at the University of Toronto Division of Orthopedics, where a set five-year training program transitioned to a CBME program after a four-year pilot.

The purpose of this thesis was to answer two primary research questions. Firstly, to determine what combination of assessment tools could be developed for use as in-training assessment of resident competency, in the setting of CBME that demonstrate evidence of validity, and are both feasible and acceptable? Secondly, to determine if it was possible for both junior and senior residents to achieve a minimal level of competence after training in CBME rotations, in order to examine the effect of clinical experience in terms of time.

In order to answer these research questions, five studies have been presented – this section will discuss the findings of each in relation to the two primary research questions.

THE ACQUISITION OF CLINICAL SKILLS

In Chapter 4, an end of rotation OSCE was used to examine residents’ clinical skills after a sports medicine rotation. These skills included history taking, examination, differential diagnosis, image interpretation, consent, clinical decision-making, and surgical techniques. Using modern theory as a framework to gather multiple sources of evidence of validity, this OSCE had evidence of content validity, with clearly detailed exam creation using a blueprint, exam review by independent content experts, rater training, and a clear rationale for using checklists and global ratings. Psychometric analysis demonstrated evidence of good internal consistency, with a Cronbach’s alpha
of 0.87 for the overall global rating scores. Analysis using G theory identified that the major source of variance was the level of training. While no relationship with other variables was sought in this study, a previous OSCE study by Dwyer et al. (Appendix C) had demonstrated a good correlation between OSCE scores and a yearly MCQ exam (OITE) undertaken by all orthopedic residents (Dwyer, Theodoropoulos et al. 2013). With regards to consequences, all results were provided to the residents, and forwarded to the program director. Table 19 demonstrates the evidence of validity that was collected for this study.
<table>
<thead>
<tr>
<th>Category of validity evidence</th>
<th>Evidence</th>
</tr>
</thead>
</table>
| Content                       | • Detailed test blueprint  
|                               | • Correlation with curriculum  
|                               | • Qualified writers  
|                               | • Content expert review |
| Response Process               | • Exam clarity – trial OSCE with residents  
|                               | • Exam clarity – good feedback on questionnaire from pilot study  
|                               | • Rater training  
|                               | • Clear rater thought process  
|                               | • Quality control of score  
|                               | • Clear score reporting  
|                               | • Rationale of GRS and checklists |
| Internal Structure             | • Excellent internal consistency (0.87)  
|                               | • Split Cronbach’s alpha good to excellent for each group (juniors, seniors and fellows) (0.7 – 0.8)  
|                               | • Good correlation between checklist and global rating scale (0.65 – 0.89)  
|                               | • No increase in Cronbach’s alpha if Station Deleted for any station  
|                               | • G Theory – 64% of variance due to level of training  
|                               | • Subgroup analysis – no effect was seen with regards familiarity save for station 1 |
| Relations to other variables  | Pilot study (Dwyer et al. 2013)  
|                               | • Significant effect of the previous number of sports rotations undertaken with the total OSCE test scores  
|                               | • No significant effect on performance between traditional training stream and competency-based training  
|                               | • High correlation with OITE total test scores and sports subscores |
| Consequences                  | • All results forwarded to program director |

Table 19. Evidence of validity collected for the clinical skills OSCE.

In this study, as well as in the EPA and OSATS study, a combination of checklists and global rating scales were used. This was done for two reasons. Firstly, there is some evidence that using this combination can help increase reliability of the scoring system. Secondly, checklists were used to help examiners to focus on the relevant components of each station. The author of this thesis believes that marking a
checklist can be useful to help to improve the concentration of examiners in stations. However, the final determination of competence was based upon global ratings, rather than checklists scores.

Using an OSCE to assess clinical skills after the sports medicine rotation proved to be a feasible assessment tool over the 18-month period, with faculty involvement limited to an hour every three months – anecdotally this was well tolerated by staff. Furthermore, the previous OSCE study had shown, using a questionnaire, that 80% of faculty agreed or strongly agreed that this was the best method of assessing medical knowledge (Dwyer, Theodoropoulos et al. 2013). Costs, aside from faculty time (which was donated), were minimal, with faculty acting as standardized patients during aspects of some stations.

The main finding of Chapter 4 was that there was a significant difference between junior and senior residents with regards the total checklist scores and overall global rating scale. Despite all residents having been exposed to an identical curriculum and teaching program, and relatively similar clinical training, significant differences in performance were also seen for each of the domains of knowledge (history-taking, consent, differential diagnosis etc.). Using a non-compensatory standard setting method, whereby residents had to pass 4/6 stations in order to pass the OSCE, only 8/21 (38%) junior residents passed, compared to 18/18 (100%) senior residents. The standard of 4/6 stations was chosen as it designates that residents must pass 65% of the examination, a consistent (albeit arbitrary) standard set by the University of Toronto and many other institutions.

Clearly, many junior residents could not achieve a minimal level of competence in clinical skills, despite training within a CBME program. Figures 1&2 display boxplots for the overall global rating scales, the total checklist scores, and the overall global rating scale for each domain of knowledge – each of these demonstrate an improvement in clinical skills over time between junior and senior residents, and a curve that appear to flatten out between senior residents and fellows (page 79-80). Certainly, in this study focusing on the acquisition of clinical skills, residents transitioned between the stages described by the Dreyfus and Dreyfus model of skill
acquisition over time, with the progression in resident performance over time mirroring Benner et al. descriptions from the field of nursing, whereby the progression from novice, to advanced beginner, and to competent practitioner took between 2.5 to 3 years (Benner 1996). The acquisition of clinical skills and possible explanations for these findings are discussed more extensively in 9.2.

There has been increasing use of an OSCE to assess resident performance in the published literature, with the majority of these studies support the findings presented in Chapter 4. In a study of anesthesiology residents performing in operating room, trauma, and resuscitation scenarios, Sidi et al. identified an association between performance and training level, with PGY-4 residents having better performance scores, reduced error rates, and increased pass rates compared to PGY-2 residents (Sidi, Gravenstein et al. 2014). This improvement in performance of clinical skills over time / years of training has also been demonstrated in studies of internal medicine residents (Pugh, Touchie et al. 2014, Pugh, Touchie et al. 2015). Using data from a mandatory but formative OSCE in 244 internal medicine graduates from PGY-1 to PGY-4 over a seven year period, it was found that scores significantly increased with each year of training (Pugh, Touchie et al. 2015). However, it must be emphasized that these studies were not conducted on residents enrolled in a CBME program, and that the clinical skills assessed were wide-ranging, as opposed to subspecialty clinical skills studied in Chapter 4.

It is clearly not acceptable to have an OSCE that over 60% of junior residents fail. One option would be to take a milestones approach as advocated by the ACGME, and perhaps use a separate OSCE written for the purpose of junior resident assessment. In the ACGME Orthopedic Milestones Project, junior residents are expected to be able to reach Level 2 - using the example of anterior cruciate ligament patient care, an OSCE station might focus on a resident’s ability to take a focused history, interpret basic imaging, prescribe non-operative treatment, and describe operative options. This is in contrast to a station written for senior residents, whereby establishing a Level 4 performance would examine a resident’s ability to recognize concomitant
injuries to other parts of the knee, interpret advanced imaging such as MRI, and discuss the management of intra-operative complications.

This approach has not been taken for three reasons. Firstly, running two separate OSCEs every three months is much less feasible, especially given that it is commonly accepted that having an OSCE with adequate reliability requires a minimum number of stations (van der Vleuten and Schuwirth 2005). Running two six-station OSCEs would also double the amount of faculty time required, and likely reduce the acceptability of the assessment tool. Finally, this author believes that early and continued exposure of residents to the standard expected is an important component of training - exposure to commonly used modern modalities such as MRI is but one example. This exposure has significant educational impact – residents can self-reflect on their deficiencies in knowledge, and faculty can provide appropriate feedback on performance.

As a result of this philosophy, one in-training OSCE was used for all residents regardless of year of training, necessitating the use of alternative standard setting methods, such as the modified Angoff method. Critically, it was felt important to determine if a test-centred method could be used to create separate cut-scores for junior and senior residents, prior to them sitting the OSCE – in this way, cut-scores could be used to identify a resident in difficulty, a resident performing significantly below that of their relevant peer group.

Using the data from the first study, Chapter 5 presents the results of comparing the modified Angoff method (a test-centred method, whereby a cut-score can be created prior to test-taking), to the BG and BLR methods of standard setting (examinee-centred methods that use examination results to set standards). The results of this study demonstrated that all three groups of judges (sports medicine surgeons, sports fellows, non sports medicine surgeons) expected senior residents to score a significantly higher percentage of checklist items than junior residents. Importantly, there was no significant difference between the cut-score set for senior residents by the modified Angoff method, and the cut-score created by the BG and BLR methods. Credibility of the modified Angoff method was determined by examining the pass rates of senior residents and fellows, none of who failed the OSCE using this method. Most
importantly, the modified Angoff standard setting method was more acceptable than the non-compensatory standard setting method used in Chapter 1, with only 1/21 (4.8%) of the junior residents failing the OSCE using the modified Angoff method, compared to 6/21 (28.6%) with the BG and BLR methods, and 8/21 (38%) with the non-compensatory standard setting method.

The BR and BG methods are commonly used in the setting of medical and dental school examinations, where the results of hundreds of students are used to create a cut-score, anchored by the results of students given a global rating of borderline (Kaufman, Mann et al. 2000, Schoonheim-Klein, Muijtjens et al. 2009). These methods are not practical in the setting of CBME. At the University of Toronto Division of Orthopedics, the largest orthopedic residency-training program in the Canada, only five to six residents undertake the sports medicine rotation at any given time. This is most likely insufficient data to conduct a BR or BG method after each iteration of the OSCE – not only are there limited data points, but by definition there will be limited numbers of residents assigned as borderline for each station. Furthermore, a test-centred method such as the modified Angoff seems ideal in the setting of CBME because a cut-score can be set prior to residents undertaking the OSCE, allowing those who are underperforming to be immediately identified. Finally, the results of Chapter 4 demonstrate that judges can use a modified Angoff method to set different cut-scores for junior and senior residents, enabling the continued use of the same OSCE to focus on identifying those who are underperforming in comparison with their peer group.

While there are studies demonstrating that the modified Angoff method is an acceptable and credible method of setting standards in undergraduate medicine (Morrison, McNally et al. 1996, Kaufman, Mann et al. 2000), the use of standard setting has not been studied extensively in postgraduate medicine. Using a 16 station OSCE to assess final year general practice trainees, Kramer et al. compared a modified Angoff method to a BR method; using a modified Angoff method with reality check, 66% of trainees passed, while using the BR method 95% passed (Kramer, Muijtjens et al. 2003). Not only were these results deemed unacceptable, this study
also found that many experienced GPs failed the OSCE using the modified Angoff, questioning its credibility. While the results of the study by Kramer et al. contrast significantly with the findings presented in Chapter 5, it is important to note that Kramer’s et al. description of a modified Angoff method related to the lack of group discussion; judges were expected to estimate the proportion of borderline trainees that would correctly perform each item. In other modifications of the Angoff method (including that used in Chapter 5), the modification is to use ‘yes/no’ in regards whether a borderline candidate would answer each question correctly (Downing, Tekian et al. 2006, Schoonheim-Klein, Muijtjens et al. 2009). This may explain the disparity between results, as predicting the proportion of borderline trainees who would answer a particular checklist question correctly is difficult. However, as there is limited use of standard setting methods in the postgraduate medical education literature, this is an area that warrants further investigation.

In the study presented in Chapter 5, a SEM was added to the pass marks established by the BG / BLR methods. The literature on the use of SEM in this manner is varied. Wilkinson et al. in 2001 described using the BG method in an OSCE assessing 5th year medical students – in this study no SEM was added (Wilkinson, Newble et al. 2001). Wood et al. in 2006 compared the BG to the BLR, identifying that there were considerably more failures with the BLR – no SEM were added to any of the cut scores (Wood, Humphrey-Murto et al. 2006). However, in a medical student OSCE, Pell et al. described adding 2 SEM to the BLR to prevent false positives (Pell, Fuller et al. 2013). In the study presented in Chapter 5, a SEM was added to all cut scores in order raise the bar, and to set the pass mark at the competent rather than borderline grade.

**Conclusion**

Using van der Vleuten’s conceptual model, an OSCE has been shown to be a valuable tool for the evaluation of clinical skills in the setting of CBME. The results of Chapter 3 demonstrate that there is sufficient evidence of validity to support its use in the setting of a medium-stakes in-training examination. Acceptability was high, as faculty were familiar with the format from their own education and training experience.
Furthermore, because high-stakes exit examinations use an OSCE format, residents welcome early and regular exposure to this assessment tool. The OSCE was also extremely feasible, requiring limited resources in terms of both time and cost. Finally, the OSCE had clear educational impact – not only were the clinical scenarios useful to residents in providing important feedback on knowledge and performance, but using the modified Angoff standard setting method allowed the identification of residents who are underperforming, critical to the underlying premise of CMBE.

THE UNDERSTANDING AND APPLICATION OF THE INTRINSIC CANMEDS ROLES

Chapter 6 focuses on the assessment of the intrinsic (non medical expert) CanMEDS roles, presenting the results of a study that used an OSCE to assess the understanding and application of these roles in three groups of residents; PGY-0, PGY-3 and PGY-5. Using a modern validity theory as a framework, the CanMEDS OSCE had evidence of content validity, with use of a detailed blueprint, case review by independent content experts, and provision of a rationale for use of global ratings. Clear test instructions were provided for the residents, rater training was provided, and data integrity was aided by the use of online scoring and electronic data collection. Evidence of internal structure was represented by a good internal consistency for the overall rating of performance, as well as for four of the six intrinsic roles. Relationships with other variables were demonstrated in two ways. Firstly, a good correlation seen between the program directors’ ranking of the residents prior to the CanMEDS OSCE, and resident performance on the OSCE. Secondly, a good correlation was also seen between the overall ITER scores of the PGY-3 and the PGY-5 residents, and the scores of the Communicator, Manager and Professional roles on the OSCE. No ITER scores were available for the PGY-0 residents, as it was at the beginning of their training. The evidence of validity collected for this study is presented in Table 20.
<table>
<thead>
<tr>
<th>Category of validity evidence</th>
<th>Evidence</th>
</tr>
</thead>
</table>
| Content                      | • Detailed test blueprint  
                                 • Correlation with curriculum  
                                 • Qualified writers  
                                 • Content expert review |
| Response Process             | • Rater training  
                                 • Clear rater thought process  
                                 • Rationale for using global rating scales  
                                 • Global rating scale descriptors  
                                 • Quality control of score  
                                 • Clear score reporting  
                                 • Exam clarity – good feedback on questionnaire |
| Internal Structure           | • Excellent internal consistency  
                                 • Excellent internal consistency for 4/6 roles  
                                 • No G Study |
| Relations to other variables| • Good correlation between the PD ratings of intrinsic roles, and the OSCE roles  
                                 • Correlation seen between ITER overall score and communicator, manager and professional role scores on the OSCE |
| Consequences                | • In the following iterations of the CanMEDS OSCE, any resident deemed to be less than competent at any station or role undergoes remediation |

**Table 20.** Evidence of validity collected for the CanMEDS OSCE.

The CanMEDS OSCE was felt to be acceptable, as determined using a questionnaire filled out by faculty and residents - nearly 90% of responding residents agreed that the scenarios reflected realistic clinical scenarios. There were some feasibility issues - faculty donated four hours of time to conduct the study, and the costs of the SP’s was not insignificant, with a total cost of nearly $3,500. It is likely that the educational impact was significant, with exposure of junior residents to difficult clinical scenarios, some before their residency began.
The CanMEDS OSCE demonstrated a significant difference in total test scores between the PGY-5 residents, and both the PGY-3 and PGY-0 residents – this finding was also seen in five of the six stations, and for each of the six CanMEDS roles examined. Despite a trend towards increased scores in the PGY-3 residents, no difference was seen between the PGY-0 (6 residents) and PGY-3 (13 residents) groups, although this study was clearly underpowered.

The second aim of this thesis was to determine if all residents could achieve a minimal level of competence. In the CanMEDS OSCE, residents were assigned an overall GRS for each station - in order to determine who would pass the CanMEDS OSCE, a non-compensatory method was applied, whereby competence or better had to be achieved at 4/6 stations (Table 21). The results of this standard setting method demonstrate once again that there are differences in the ability to understand and apply the intrinsic CanMEDS roles based upon year of training.

<table>
<thead>
<tr>
<th>Stations</th>
<th>PGY 0</th>
<th>PGY3</th>
<th>PGY5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Station 1: Needlestick</strong></td>
<td>4/6 (67%)</td>
<td>12/13 (95%)</td>
<td>6/6 (100%)</td>
</tr>
<tr>
<td>Roles: Professional, Manager</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Station 2: Trauma List</strong></td>
<td>0/6 (0%)</td>
<td>9/13 (69%)</td>
<td>5/6 (83%)</td>
</tr>
<tr>
<td>Roles: Manager, Communicator</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Station 3: Non accidental Injury</strong></td>
<td>6/6 (100%)</td>
<td>13/13 (100%)</td>
<td>6/6 (100%)</td>
</tr>
<tr>
<td>Roles: Health Advocate, Manager</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Station 4: Breaking bad news</strong></td>
<td>5/6 (83%)</td>
<td>12/13 (95%)</td>
<td>6/6 (100%)</td>
</tr>
<tr>
<td>Roles: Communicator, Professional</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Station 5: Interacting with OR team</strong></td>
<td>2/6 (33%)</td>
<td>9/13 (69%)</td>
<td>6/6 (100%)</td>
</tr>
<tr>
<td>Role: Collaborator</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Station 6: Spinal Evidence</strong></td>
<td>2/6 (33%)</td>
<td>11/13 (84%)</td>
<td>6/6 (100%)</td>
</tr>
<tr>
<td>Role: Scholar, Communicator</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 21. Results of the CanMEDS study, demonstrating the number of residents in each group able to achieve a minimal level of competence on 4/6 stations.
These findings are not necessarily intuitive – many of the faculty at the University of Toronto initially postulated that some of the skills being examined would relate to the personality traits or previous life experiences of residents. For this reason, many of the faculty believed that incoming residents would be able to perform as well in the clinical scenarios as more senior residents. After the OSCE, it seemed evident that for many residents, previous experience and exposure to difficult clinical situations in their training (cancelled patients, delivering bad news) had provided them with important skills useful in managing these scenarios.

The OSCE format is being increasingly used to assess the intrinsic CanMEDS roles in residents (Nguyen, Tardioli et al. 2015) - for the most part, the literature supports the finding that experience and level of training is associated with improved performance (Keely, Myers et al. 2002, Jefferies, Simmons et al. 2007, Nicksa, Anderson et al. 2015). Recently, Nicksa et al. reported on using simulation to assess surgical residents’ demonstration of leadership skills, teamwork, and effective communication – a significant difference was seen in performance of the PGY-2 residents compared to the PGY-1 residents (Nicksa, Anderson et al. 2015). Furthermore, over two testing periods using debriefings and feedback, it was found that the PGY-2 residents’ performance improved, while that of the PGY-1 residents did not. Nicksa et al. postulated that these findings might be the result of the steep learning curve experienced by first year residents (Nicksa, Anderson et al. 2015). Pugh et al., studying general medicine residents, combined an assessment of technical skill performance with performance in three of the intrinsic CanMEDS roles; Professional, Communicator, and Collaborator - in this study, PGY-3 residents outperformed PGY-1 residents in the so-called non-technical components of the OSCE (Pugh, Hamstra et al. 2015).

Not all studies concur with the findings presented in Chapter 6. Falcone et al. used a two station OSCE to assess the ability of junior and senior surgical residents to hold difficult conversations; junior residents discussed the risks and benefits of surgery, and informed patients of a cancer diagnosis, while senior residents explained goals of care to a terminal patient with a surgical emergency, and managed a transition to palliative
care (Falcone, Claxton et al. 2014). In this study, standardized patients judged that junior residents had a higher degree of communication skills than senior residents - however it was acknowledged by the authors that having the two groups of residents participate in different scenarios likely influenced these findings. Nguyen et al. used a single station to evaluate the technical and non-technical skills of urology residents (PGY3-5) – while this station was unable to distinguish between residents of different years of training, only nine residents participated, and little evidence of validity was provided (Nguyen, Tardioli et al. 2015).

A recent survey of program directors in Canada identified that the most common assessment used for the role of Professional was feedback from faculty, and other health professionals, typically in the form of an ITER and / or a 360-degree evaluation (Warren, Allen et al. 2014). At the University of Toronto, the ITER continues to be used, despite the many limitations discussed in the literature review. Interestingly, while the results of the study presented in Chapter 6 revealed a correlation between resident’s overall ITER score, and the roles of Communicator, Manager, and Professional, the only correlation between specific roles on the ITER and specific roles on the OSCE was the role of Communicator. Furthermore, there was a good correlation between the OSCE scores and the PD rankings. This suggests that while the ITER can provide some guidance to resident performance, it has a limited ability to assess performance across the six intrinsic roles. This is supported by some of the evidence around ITERs that is presented in the literature review.

The main advantage of the OSCE is the creation of simulated scenarios, whereby residents are placed into uncomfortable and / or difficult environments, and given a safe opportunity to participate without risk to patients. Junior residents are able to experience unusual situations for the first time, obtain direct feedback on performance, and prepare for similar situations in the future (Nicksa, Anderson et al. 2015). The OSCE is now being used on an annual basis, to assess PGY-3 residents and identify and remediate any perceived deficiencies in any of the roles.

While the study in Chapter 6 provides a mechanism for the assessment of the intrinsic roles, and information on how resident’s progress over time in their ability to apply
these roles, how to assess the intrinsic roles after or during rotations in CBME remains an issue. The curriculum map at the University of Toronto Division of Orthopedics (Appendix B) designates specific roles to be both taught and assessed during each of the 21 rotations, with separate roles for juniors and seniors. In the sports rotation for example, the Communicator and Professional roles are assigned to the junior rotation, while the roles of Scholar and Professional are assigned to the senior rotation. At this time the ITER continues to be used as the assessment tool – another possibility would be to incorporate CanMEDS stations into the end-of-rotation clinical skills OSCE, a model used in the orthopedic certification examination held by the RCPSC. This approach has now being used in the end-of-rotation OSCE for the arthroplasty rotation. In this OSCE, the stations focus on the blueprinted roles for junior and senior residents, but are being used primarily for formative reasons.

**THE PERFORMANCE OF TECHNICAL PROCEDURES**

Chapter 7 presents the results of an OSATS, used to assess the performance of technical procedures by residents after a sports medicine rotation. Once again, there was evidence of content validity - the exam was carefully mapped to the curriculum, a detailed explanation for exam creation was provided, and expert review occurred. Clear instructions were provided to residents, and raters had significant experience in assessing residents, having participated in previous studies that evaluated the use of the shoulder and knee models to assess resident performance. The use of electronic scoring again ensured quality control of data. Internal structure was demonstrated in two ways. The internal consistency was excellent, with a Cronbach’s alpha of 0.9 - there was also a strong correlation between examiner scores and blinded review of videotaped procedures. Further analysis using G theory demonstrated an overall reliability of 0.78, with the major source of variance identified as the training level of the participant. In regards relationship with other variables, a good correlation was seen between performance on the end of rotation clinical skills OSCE, and the total ASSET score on the OSATS. All results were provided to the residents, and forwarded to the program director. The evidence of validity collected for this study is detailed in Table 22.
### Table 22. Evidence of validity collected for the OSATS study.

The OSATS was certainly feasible in terms of time – this exam format took an hour of time per resident. The cost per resident was not insignificant, calculated to be $185 per resident. The OSATS was acceptable to faculty – residents were being asked for the first time to demonstrate their ability to perform technical procedures independently. Finally, there was significant educational impact – many deficiencies were identified in the ability of residents, which has focused attention on the surgical training in the rotation, and stimulated many questions, which are discussed below.
Once again, a significant difference was seen in performance between junior and senior residents, with a significant difference detected in the overall GRS, the total ASSET scores and the total checklist scores between the two groups. Less than 50% of junior residents were able to achieve a minimal level of competency on any of the technical procedures, despite having been trained in these procedures in the operating room and on the models. **Figures 3 and 4 in Chapter 7 (p145)** demonstrate a progression over time in training / experience in the ability to perform of technical procedures, between junior residents, senior residents, fellows, and staff surgeons. In order to determine the number of participants in each group who were able to pass the OSATS, a non-compensatory method has been applied, whereby a global rating of competent or better had to be achieved in 4/6 stations (Table 23).

<table>
<thead>
<tr>
<th>Junior</th>
<th>Senior</th>
<th>Fellow</th>
<th>Staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/19</td>
<td>7/8</td>
<td>7/7</td>
<td>7/7</td>
</tr>
<tr>
<td>(10.5%)</td>
<td>(87.5%)</td>
<td>(100%)</td>
<td>(100%)</td>
</tr>
</tbody>
</table>

**Table 23.** Results of the OSATS study, demonstrating the number of participants in each group able to achieve a minimal level of competence on 4/6 stations.

In contrast to the clinical skills OSCE, the OSATS study demonstrated that senior residents also have difficulty in achieving a minimal level of competence in some of the technical procedures, in contrast to the sports medicine fellows and faculty surgeons. **Table 3 in Chapter 4 (p95)** presents the results of stations focusing on knee procedures - both senior and junior residents had difficulty drilling a femoral tunnel using an anteromedial approach. However, if this station is excluded, between 80 and 100% of senior residents were able to achieve a minimal level of competency for the remainder of the knee stations, while only between 12 and 44% of the junior were able to do so. Performance on the shoulder stations was worse (**Table 4, p96**) - while staff and fellows were able to perform these shoulder procedures competently in the simulated setting, with the exception of insertion of rotation cuff anchors,
significant numbers of both junior and senior residents had difficulty performing these procedures.

The finding that junior and senior residents cannot perform some of the sports medicine procedures in the simulated setting warrants attention. This finding is task dependent – while some tasks are more difficult for the junior residents, some procedures are too difficult for many of both groups. The latter is especially seen with arthroscopic shoulder procedures - it may be that these tasks require advanced skills, acquired during fellowship training and after significant experience. It may also highlight deficiencies in the training program - certainly in the author’s experience, in the operating room staff surgeons are reluctant to allow junior or senior residents to perform these procedures due to the complexity of the individual steps, and the high risk of complications should particular steps be performed incorrectly. Recent research is also emerging that, even with relatively simple procedures such as knee arthroscopy, there is a minimum number of procedures that must be performed prior to residents becoming competent (Hodgins, Veillette et al. 2014) – it seems likely that insufficient operating room experience with these procedures is at least partly to blame.

How to best deal with these findings is unknown. One approach would be to increase the frequency of teaching in the simulation laboratory - however these are already being held twice month over the three-month rotation. These training sessions are not without cost, both in terms of faculty time, and costs of models and implants – for this reason the number of skills sessions cannot be easily increased. Alternatively, the findings of Chapter 7 can be used to change the focus of the end of rotation assessment, by implementing a milestones approach. In this way, while all teaching and laboratory training sessions would continue to cover all of the surgical procedures, junior residents would only be expected to demonstrate a minimal level of competence at those tasks (partial meniscectomy, whipstitch hamstring graft), that the study presented in Chapter 7 indicates they are capable of performing. Similarly for senior residents, where performance of ACL reconstruction and rotator cuff surgery should
be the focus of the assessment, as it is known that the majority of senior residents should be able to perform these procedures competently.

Clearly, it will be critical to determine if these findings will be replicated in the operating room. While there is no reason to believe that residents who cannot perform these procedures in the simulated setting will be able to in the workplace, this is an important area of future research in CBME. However, there is some emerging evidence to support the findings of Chapter 7. In 2014, Hopmans et al. reported using an OSATS in the operating room to study performance of a variety of surgical procedures – scores tended to improve with year of training for the majority of procedures (Hopmans, den Hoed et al. 2014). Furthermore, this finding was also procedure specific - mid level residents were able to master relatively simple procedures, while more complex laparoscopic procedures were mastered more gradually. These findings are consistent with the results of the sports medicine OSATS study, and provide further support for a milestones based approach to CBME.

In the both the ACLR sawbones study (Appendix D), and the OSATS study, the determination of competency was done so based on an overall GRS anchored to the Dreyfus and Dreyfus model of skill acquisition, although both a checklist and the ASSET global rating scale were also used. In the original publication of the ASSET GRS, the authors reported that a minimum score of three in each of the eight domains was required to assign a grade of competent (Koehler, Amsdell et al. 2013). In the ACLR study, no significant difference in pass rates was seen depending on whether the overall GRS, or the ASSET score was used to determine a minimal level of competence. This is important, as it is unclear whether the final determination of competence should be based upon a global rating score, or the overall judgment of performance by a senior surgeon. Regardless, both of these methods are affected by a degree of subjectivity, which cannot be removed from assessment tools, only minimized.

While standard setting methods were successfully used to set cut-scores in the clinical skills OSCE, there are few reports in the literature regarding the application of standard setting methods to the performance of technical procedures. In 2013, Cohen
et al. established a minimal passing score for the insertion of central venous catheters, identifying that regular review of performance is required to ensure standards are fair and rigorous (Cohen, Barsuk et al. 2013). The cumulative summation test for learning curve (LC-CUSUM) is a statistical method that has been developed to monitor progression towards competency, over a series of procedures, by determining when the learning curve for a procedure is complete (Biau and Porcher 2010). This has been used to good effect in both surgeons and trainees to detect when a flattening trend in performance occurs – allowing the determination of when a trainee or surgeon becomes competent, as well as the estimation of the number of procedures required to achieve a minimal level of competency (Jowell, Baillie et al. 1996, Salowi, Choong et al. 2010, Miskovic, Wyles et al. 2011). However, this type of analysis requires a series of data points, and cannot be used for assessment of a single procedure. For this reason, assigning critical competencies and procedures for trainees based upon their level of training may be the best approach.

The use of simulation to assess the performance of technical procedures in residents continues to increase, with the majority of these studies demonstrating progression of psychomotor skills and performance in association with year of training and experience. These findings have been identified in procedures in the surgical specialties of urology (Argun, Chrouser et al. 2015), pediatric orthopedics (Moktar, Popkin et al. 2014, Bradley, Moktar et al. 2015), neurosurgery (Ghobrial, Balsara et al. 2015), and general surgery (Husslein, Shirreff et al. 2015). Again, it is important to note that these studies are not being held in the setting of CBME, so it is not possible to conclude whether residents of different experiences and level of training would be able achieve a minimal level of competence after a specific subspecialty rotation. However, the results of these studies support the findings presented in Chapter 7, that performance of technical procedures progresses over time and experience.

The OSATS study sought and identified a correlation between performance of technical procedures, and the OSCE scores for each resident. However, it must be acknowledged that the findings of studies seeking a correlation between medical knowledge and technical skills has been mixed (Bann, Khan et al. 2004). A recent
study looked at simulated surgery on a distal radius fracture, and failed to identify a
correlation between knowledge testing, and biomechanical testing of the surgical
construct (Putnam, Kinnucan et al. 2015). In 2009, Van Heest et al. reported on the
assessment of orthopedic residents performing a carpal tunnel release on a cadaver,
in association with a knowledge test based upon surgical anatomy, indications, and
steps – while a knowledge score < 70/100 predicted failure on the OSATS, a score of
> 70/100 did not ensure that the examinee would pass the test (Van Heest, Putnam et
al. 2009). The findings of the latter study are plausible, - clearly there is a threshold of
knowledge that is required in order to perform a technical procedure. Above that
threshold, a combination of innate technical abilities and surgical experience has an
important role to play in the ability to perform a particular procedure to a competent
level.

A recent systematic review by Hatala et al. concluded that there was sufficient
evidence for the use of OSATS in formative feedback and program evaluation, but
less so for high-stakes assessment, due to an absence of linkage to performance in
real clinical settings (Hatala, Cook et al. 2015). While performance in in the clinical
setting will remain an important area of future research, the use of OSATS as an in-
training assessment using simulation has shown evidence of sufficient validity for
continued use in these medium-stakes examinations. Future iterations of the OSATS
in the CBME format should focus on those procedures that junior and senior residents
can realistically achieve competence at, while providing experience, exposure and
appropriate feedback for the performance of other tasks.

ENTRUSTABLE PROFESSIONAL ACTIVITIES

Chapter 8 presents the results of a study examining simulated assessment of EPAs by
orthopedic residents, focusing on pre- and postoperative management (OSCE format),
and the performance of technical procedures (OSATS). Mapped to the top 10 EPA’s
as designated by faculty, PGY-1 and PGY-4 residents were assessed on their ability
to perform three EPAs; managing a patient with hip fracture, ankle fracture, and a
patient for a total knee replacement. All of the PGY-1 residents involved in this study
were at the end of the first year of training, and each had been deemed competent in
the CBME rotations of basic trauma, hip fracture, and basic arthroplasty. The performance of the PGY-1 residents was compared to PGY-4 residents, all of whom had recently finished complex trauma and arthroplasty rotations.

With regards to evidence of validity, independent experts reviewed all content, clear instructions were provided to residents, and electronic scoring was used to ensure data quality. Once again, all examiners had significant experience in rating residents in both the OSCE and OSATS formats. Similar to the OSATS study, all checklists were created by way of a modified Delphi procedure, involving sports medicine surgeons, arthroplasty surgeons, trauma surgeons as well as the program and assistant program directors. Three rounds of structured questionnaires were used, with the first based upon review of the curriculum and the literature. In the survey, members were asked to rate each checklist from 1 to 10 – items that were rated 8/10 or higher by 7/8 surgeons were accepted, while items rated < 4/10 by 75% of surgeons were removed. This process was repeated until consensus was achieved.

A high reliability was seen for each of the three EPAs (hip fracture, ankle fracture, and total knee arthroplasty); there was also a strong inter-rater reliability for the 6/9 stations that were videotaped and reviewed. While not published in the manuscript due to the editor’s request, all residents and examiners filled out a written questionnaire, regarding the acceptability of this assessment tool. On the questionnaire, 17/18 (94%) of residents and 9/9 (100%) of examiners agreed or strongly agreed that the scenarios reflected those that would be encountered in a general orthopedic practice, while 15/18 (83%) residents agreed or strongly agreed that the simulation was an accurate reflection of real practice. Overall, 9/9 (100%) of examiners thought that an EPA exam was a good approach to assessing a resident’s required level of supervision when managing patients. All residents received a written summary of their results for each station, and all results were forwarded to the program director. The evidence for validity for the EPA study is presented in Table 24.
<table>
<thead>
<tr>
<th>Category of validity evidence</th>
<th>Evidence</th>
</tr>
</thead>
</table>
| Content                      | • Detailed test blueprint  
|                               | • Correlation with curriculum  
|                               | • Qualified writers  
|                               | • Creation of checklists using modified Delphi procedures  
|                               | • Content expert review |
| Response Process              | • Rater training  
|                               | • Clear rater thought process  
|                               | • Rationale for using global rating scale and checklists  
|                               | • Quality control of score  
|                               | • Clear score reporting |
| Internal Structure            | • Excellent internal consistency (0.84-0.89)  
|                               | • Good to high correlation between overall checklist scores and overall global rating (0.71)  
|                               | • No increase in Cronbach’s Alpha if Item Deleted for any station  
|                               | • Excellent inter-rater reliability (0.87)  
|                               | • No G study |
| Relations to other variables  | • Poor relationship to previous performance on related rotations |
| Consequences                 | • Minimal |

Table 24. Evidence of validity collected for the EPA study.

Feasibility was an issue with this assessment method. The testing of resident performance of three EPAs took a day out of clinical practice for six faculty and a half day for 18 residents. Costs for the EPA study were also significant - while not reported in the manuscript, costs averaged $270 per resident, which does not account for faculty time, implants or instrumentation. Whether this study will be repeated on a yearly basis is unknown at this time.

The EPA study demonstrated a significant difference in performance, with regards the mean global rating scale, between junior and senior residents on 7/9 stations of the three EPAs - the preoperative management of the hip fracture patient, the postoperative management of the knee replacement and ankle fracture patient, as well as all three technical procedure stations. The results of the simulated technical
procedures were especially striking - 6/9 PGY-1 residents failed to achieve a minimal level of competence on the hip fracture procedure, 6/9 failed the arthroplasty procedure, and 7/9 failed the ankle fracture procedure. These findings are concerning, given that each of the PGY-1 residents had been previously been deemed competent after basic trauma, hip fracture and arthroplasty rotations in the CBME program. Table 25 details the number of residents in each group able to achieve a minimal level of competence on all three stations of each EPA.

<table>
<thead>
<tr>
<th>EPA</th>
<th>PGY-1</th>
<th>PG-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ankle fracture</td>
<td>1/9 (11%)</td>
<td>4/9 (44%)</td>
</tr>
<tr>
<td>Hip fracture</td>
<td>0/9 (0%)</td>
<td>7/9 (78%)</td>
</tr>
<tr>
<td>TKA patient</td>
<td>3/9 (33%)</td>
<td>7/9 (78%)</td>
</tr>
</tbody>
</table>

Table 25. The number of residents in each group able to achieve a minimal level of competence in all three stations of each EPA.

There are a few possible explanations for the findings that junior residents were not able to perform these relatively simple technical procedures competently. The first possibility is that, without intra-operative correlation, the technical procedure stations do not adequately represent the operative procedure. Against this theory is the finding that the majority of senior residents were able to perform these procedures competently. It is also unlikely that a resident who cannot perform this procedure safely and independently in the simulated setting will be able to in the operating room.

Another, and more likely, explanation is that in the current CBME program all junior residents are being assessed on their performance of technical procedures in the operating room. The procedures assessed vary from resident to resident, and not all will have been assessed on performance of hip and ankle fracture fixation for
example. Furthermore, in setting of an end of rotation assessment of operative performance, faculty cannot allow procedures to be performed below standard, as good clinical outcomes and patient safety must be ensured at all times. Finally, the assessment is performed by the faculty supervisor, which introduces many confounders. Once again, this is the advantage of using simulation – residents are able to perform tasks completely independently, without risk to patient safety, and are assessed by multiple faculty, enabling deficiencies in ability and skill to be identified.

Some junior residents also had difficulty passing the pre- and postoperative management stations of each EPA, despite having been deemed competent in the CBME rotations. A possible explanation is that the assessment tools being used in the rotations other than sports medicine are not sufficiently robust to highlight deficiencies in clinical skills. At this time, in rotations such as trauma and arthroplasty, assessment takes the form of structured oral examinations and / or long cases, with the faculty preceptor acting as the examiner – these assessment tools have many limitations as discussed in the literature review. Another explanation for this finding is revealed by close inspection of the ACGME Orthopedic Milestones Project. For example, in this document the diagnosis and management of complications (covered in each postoperative management station of the EPA study) is a component of Level 4 resident performance, a resident who is ready to graduate. The management of complications generally requires high levels of knowledge, as well as some previous experience with these difficult and uncommon outcomes, potentially explaining the difficulty some junior residents had with the postoperative management stations.

It is interesting to note that, although less frequently than the junior residents, some of the senior residents were not able to achieve a minimal level of competence, in both the clinical skills and technical procedure stations. Certainly, this highlights the benefits of objective assessment in CBME – each of these EPAs have been highlighted as a critical element of orthopedic training, thus mandating that all residents demonstrate a minimal level of competence in each prior to graduation. The identification of those few senior residents who require remediation allows additional training to be provided - prior to the development and evaluation of robust assessment
tools in a CBME program, it is likely that these deficiencies would not have been identified, as there is no assessment of performance of technical procedures at the RCPSC orthopedic certification examination. In this way, objective assessment of residents’ performance of EPA’s prior to graduation can provide a new level of accountability, an important element of an outcomes-based program.

At this time, the majority of published research surrounding EPAs involves the development of lists of critical EPAs for specialties and subspecialties (Chan, Englander et al. 2014, Fessler, Addrizzo-Harris et al. 2014, Rose, Fix et al. 2014, Myers, Krueger et al. 2015, Schultz, Griffiths et al. 2015). The continued volume of literature surrounding EPAs speaks to the popularity of the concept, as EPAs can be used to divide a specialty curriculum into critical competencies, with a direct link to clinical practice (Yuan, Prince et al. 2014). This use of EPAs as an organizing framework of assessment has been described as potentially superior to the ACGME general competencies (Caverzagie, Cooney et al. 2015). Furthermore, assessment tools can be developed that focus on EPAs, or those tasks deemed the most important for residents in training programs to perform competently.

Some literature has emerged reporting the assessment of medical graduates’ ability to perform EPAs. Recently, Yuan et al. reported on the use of chart audits to assess nephrology fellows’ performance in managing a nephrology outpatient clinic – many deficiencies were identified that required remediation (Yuan, Prince et al. 2014). In 2015, Shumway et al. evaluated hematology / oncology fellows’ ability to perform five EPAs independently – examples included writing chemotherapy orders, managing toxicities and managing end-of-life issues (Shumway, Dacus et al. 2015). Using faculty discussion to determine competence, a progression towards competency was seen over time, as fellows progressed from PGY-4 to PGY-6. Finally, Shaughnessy et al. used an assessment form to report on residents’ performance of 60 EPAs in family medicine – competence was statistically different among residents of different years of training, and increased with resident experience over time (Shaughnessy, Chang et al. 2014). It was interesting to note that no resident, over the 14-month collection period for this study, was able to complete an assessment form for all 60 EPAs, lending some
credence to the notion that no more than 20 – 30 EPAs are manageable for most graduate medical education programs (Ten Cate 2013).

**SUMMARY**

In proposing a conceptual method for defining the utility of an assessment method, van der Vleuten placed emphasis not just on the validity and reliability of an assessment tool, but also on important issues of feasibility, accountability, and educational impact. While this model is not designed for use as an algorithm, compromise can occur in certain components depending on the context, proposed use and resources available. This conceptual method is not just useful in evaluating an assessment tool, but can be used to devise and evaluate an overall assessment program.

The first aim of this thesis was to develop and evaluate a program of in-training assessment tools, covering the breadth of competencies. Four of the five studies presented demonstrate evidence of validity, to varying degrees. As discussed in the literature review, validity is not dichotomous – using modern theories of validity, we do not say an instrument is valid or not valid, but rather it has less or more evidence of validity (Downing, Lieska et al. 2003). Furthermore, the degree of validity evidence required varies with the use of the instrument, with more validity evidence required for higher stakes examination (Cook, Beckman et al. 2010). While the setting of in-training CBME assessment most likely can be described as medium-stakes, it is important to ensure that these assessment tools are held to the highest standard possible due to their significant educational impact - failing end-of-rotation exams requires rotations to be repeated, potentially lengthening training. There is also a possibility that, in the setting of residents who cannot achieve a minimal level of competency after rotations, residents may have to leave training programs – for this reason the assessment tools and their findings must be defendable.

A second, important aspect of the first aim was to develop an assessment program that was feasible and acceptable to both residents and faculty. This thesis avoided the development of new assessment tools, using commonly accepted and well-
researched methods such as OSCEs and OSATS. The clinical skill OSCE was both feasible and acceptable, requiring only an hour of faculty time every three months, and having no actual costs apart from faculty time. The CanMEDS OSCE was more expensive, as many of the scenarios were dependent upon SPs - the average cost per resident was $140 per. For this reason, this assessment tool is now being used once a year, providing valuable formative feedback to residents.

There is no doubting that assessing performance of technical procedures is resource heavy. Despite using low fidelity sawbone models, the estimated cost for the OSATS study was $185 per resident, which does not take into account the costs of models used in training, nor the multitude of implanted devices (which were donated by industry). Similarly, costs for the EPA study, while not reported in the manuscript, averaged $270 per resident – once again this calculated cost does not account for faculty time, implants and instrumentation. Not all programs will have access to these funds, or to industry support. However, using high-fidelity models such as cadavers will usually cost nearly 10-fold more, while virtual reality has not advanced to the point where more complex and involved arthroscopic procedures can be performed.

Feedback from the residents and faculty was generally positive with regards the assessment tools, as demonstrated by the responses on the questionnaires regarding the clinical skills OSCE, the CanMEDS OSCE, and the EPA study. While the high failure rate of junior residents encountered in the clinical skills OSCE was clearly not acceptable, using a modified Angoff method enabled a separate standard to be set for the junior residents, resulting in a much more acceptable failure rate. While questionnaires were not used in the OSATS study, both residents and faculty seemed very accepting of using models to both practice and assess the performance of technical procedures. In fact the residents were very appreciative of any opportunity to perform procedures to which they would usually have limited clinical exposure.

The educational impact of the studies presented in this thesis was significant. The clinical skills OSCE has given the program a feasible method of assessing these important skills in an objective manner, identifying the resident in difficulty, and preparing residents for their certification examination. The CanMEDS OSCE is now a
formal component of the curriculum, exposing residents to difficult yet important clinical scenarios, and providing them with formal feedback on performance from staff. The OSATS study has given the program and the orthopedic community important information regarding which sports medicine procedures residents can perform, and provided some evidence to support to a milestones approach to training. Finally, the EPA study has emphasized the importance of creating lists of critical tasks for specialty training, and the importance of using assessment tools to ensure a minimal level of competence is being achieved prior to graduation.

The second aim of this thesis was to determine if it was possible for all residents to achieve a minimal level of competence after a rotation within a CMBE model of training – this important issue is discussed in depth in the next section.

### 9.2 IMPLICATIONS OF THESE FINDINGS IN RELATION TO CBME

#### THE ATTAINMENT OF COMPETENCE

The second aim of this thesis was, using the results of the assessment tools, to examine the premise that all residents, in the setting of CBME, would be able to achieve a minimal level of competency after a subspecialty rotation. Four of the five main papers of this thesis demonstrated a single finding; that junior residents, overall, cannot perform as well as senior residents, after training within a CBME model of education. This was evident from the two papers assessing clinical skills and the performance of technical procedures after a sports medicine rotation. A similar finding was encountered when assessing performance of EPAs, and the application and understanding of the intrinsic CanMEDS roles.

The results of these studies would suggest that there is an effect of clinical experience or duration of training that has a significant impact on residents’ ability to perform clinical skills or technical procedures. While a CBME model of training might expect that all residents, regardless of year of training, demonstrate a minimal level of competency prior to exiting a rotation, this has not been the case. This is not to say the CBME has not improved training (although this thesis is not able to confirm or
deny such a statement), but only to conclude that there remain factors outside of CBME that impact residents’ ability to acquire skills and become competent at tasks. By studying the performance of residents in this thesis, a picture has emerged of an upward trajectory of performance over time, across a breadth of competencies.

Why is there is such a distinction between junior and senior residents with regards to performance? The literature review presented theories of clinical reasoning that support these findings, and provide some explanation of how this progression might occur. Schmidt et al. described a stage theory of clinical reasoning, with students progressing through transitory stages characterized by changes in knowledge structure (higher order thinking and pattern recognition), developing from continued exposure to patients (Schmidt, Norman et al. 1990). Stage 3 becomes characterized by the use of illness scripts allowing pattern recognition of signs, symptoms and diagnoses (non analytical), while Stage 4 is characterized by the application of these illness scripts to unusual or previously unseen clinical problems (Carraccio, Wolfsthal et al. 2002). The latter two stages help describe the competent and proficient medical practitioner. Eva described two types of clinical reasoning: the non analytic method, where the learner uses pattern recognition generated from experience, and the analytic method, where there is a hypothetico-deductive approach to problem solving, helping to differentiate between the competent practitioner, and one with increasing levels of expertise (Eva 2004).

In the study using an OSCE to examine the acquisition of clinical skills, residents progressed through the stages described by Dreyfus and Dreyfus model of skill over time. This novice to master rubric was originally applied to skills such as flying a plane, driving a car and playing chess, but has been adopted by medical educators (Carraccio, Benson et al. 2008). Carraccio et al. provide descriptors for each of the stages of the Dreyfus model, integrating Schmidt’s stages of clinical reasoning. For example, the competent resident is able to use patterns of illness from previous encounters to manage many common patterns of illness, and is beginning to use these scripts to reason through unusual problems, with varying success (Carraccio, Benson et al. 2008).
While the descriptors provided by Carraccio et al., which link Schmidt’s theory of clinical reasoning to the Dreyfus model of skill acquisition, were not provided to examiners in the OSCE study, resident progression through these stages in association with year of training was seen. The majority of the cases presented in the clinical skills OSCE were usual, and common place – a competent resident should have had sufficient clinical experience to being able to recognize these common symptoms and signs, use imaging to confirm a diagnosis, and manage each clinical problem appropriately. Some stations had elements of more complex problems, such as recognizing and managing complications – these scenarios require advanced clinical reasoning, synonymous with the proficient practitioner (Carraccio, Benson et al. 2008).

The clinical skills OSCE was not designed to highlight the proficient or expert practitioner, but rather to identify the competent resident, one who could easily identify and manage common clinical problems competently. Only one descriptor for the overall global rating scale based upon the Dreyfus model was provided - examiners were instructed to designate a resident as competent if their performance would ensure a pass at the orthopedic certifying exam – a competent general orthopedic surgeon. It was clear that faculty understood this descriptor – all faculty have undertaken and passed this exam, and some have acted as examiners. While the lack of descriptors for each GRS could be criticized, it has been shown that descriptors have limitations, and that using standardized and realistic vignettes may improve discrimination of performance (Regehr 2007).

The progression of resident performance in clinical skills over time mirrored the findings by Benner et al. in the field of nursing, whereby the progression from novice to competent practitioner took between 2.5 to 3 years (Benner 1996). In the OSCE study, less than half of the junior residents (between PGY-1 and PGY-3) achieved a minimal level of competence, while all senior residents (PGY-4&5) became competent. While defining an average length of time to achieve competence was beyond the scope of the paper presented in Chapter 4, it would seem that the transition towards competence in sports medicine occurred over a similar time frame. Importantly, all
senior residents were able to achieve a minimal level of competence by their fourth and fifth years of training, and in fact their performance on this OSCE was not easily distinguished from the performance of the fellows who undertook the OSCE. While this study was underpowered to detect a difference between senior residents and fellows (and this was not a primary aim of this study), it seems reasonable to surmise that the CBME training model is working well for the senior residents.

A similar progression towards competence over time was encountered in the performance of technical procedures in the OSATS study. Fitts and Posner, proposing their theory for the acquisition of psychomotor skills, described a series of three stages (Fitts 1967). Beginning with planning steps to accomplish a task (cognitive stage), followed by integration of this knowledge into motor behavior through practice and repetition (associative stage), in the third stage routine performance no longer requires cognitive input (autonomous stage). Once again, this theory of acquisition of psychomotor skills predicts a progression in ability over time that was seen in the OSATS study. Again, while there was no attempt to link these descriptors and stages to the Dreyfus model, it was clear that those residents unable to perform procedures competently did not understand the steps required to accomplish each task (cognitive stage), and lacked the practice and repetition required to perform these steps smoothly and competently (associative stage).

Whether the differences detected in the performance of technical procedures between junior and senior residents was related to differences in exposure to sports medicine procedures, the length of time spent in sports medicine, or related to the psychomotor skills they have acquired from other rotations is unknown. The curves demonstrated in the OSATS studies mirror learning curve studies, which are common in the procedural based specialties. Many complex orthopedic procedures have known learning curves, including hip resurfacing, hip arthroscopy, and shoulder replacement (Aulakh, Jayasekera et al. 2014, Groh and Groh 2014, Hoppe, de Sa et al. 2014).

Learning curves have also been demonstrated in procedures such as ACL reconstruction (Hohmann, Bryant et al. 2010), shoulder arthroscopy (Howells, Auplish et al. 2009), and arthroscopic rotator cuff repair (Guttmann, Graham et al. 2005),
procedures performed by the residents in the OSATS study. While these studies focus on performance by fully trained orthopedic surgeons, there is some evidence that resident performance follows a similar pathway (Hodgins, Veillette et al. 2014). Each of these studies demonstrate a progression towards competence and proficiency over time, showing that outcomes such as time to completion and complication rates are linked to the number of times a task has been performed (Tekkis, Senagore et al. 2005, Hoppe, de Sa et al. 2014, Linder, Anand et al. 2015, Malas, Saczkowski et al. 2015). Whether the variable is a countable repetition, experience or time depends on the context (Pusic, Boutis et al. 2015).

In the OSATS study, time was used as the variable, with residents grouped by year of training. Other measures such as time spent practicing these procedures in the simulated setting, and the number of procedures observed or performed in the operating room was not available, and might have provided some insight into whether direct exposure to sports medicine procedures or the effect of overall clinical experience was the most important factor. However, these variables can be hard to measure, and suffer from confounders such as recall bias. What can be said is that overall, many junior residents and some senior residents were unable to achieve a minimal level of competence in some of these technical procedures, procedures that fellows and faculty were able to perform.

Research in other specialties has generated similar findings. In 2015, Pugh et al. used an OSCE to assess the procedural skills of internal medicine residents, demonstrating that PGY-3 residents scored significantly higher than PGY-1 residents on stations including lumbar puncture, endotracheal intubation and central venous catheter insertion (Pugh, Hamstra et al. 2015). In this study, the authors identified an association between the number of procedures performed, and the attainment of competency, leading them to postulate that trainees need more experience with procedures in order to become proficient, not just more time in training. However, it seems likely that performance of technical procedures is related to the number of specific procedures performed, as well as to an overall progression in psychomotor skills that occurs over time, and across all subspecialty rotations.
The data from the research presented in this thesis suggests that reaching an autonomous stage, whereby performance is routine, is task dependent (Dunphy and Williamson 2004). For example, while senior residents were mostly able to achieve a minimal level of competency in basic tasks such as ankle and hip fracture fixation, and basic knee arthroplasty (EPA study), a longer time period in association with focused training is clearly required for the majority of the sports medicine procedures in the OSATS study, where only fellows or staff surgeons were consistently able to achieve a minimal level of competency. There is support for these findings from the field of neurosurgery. In 2015, Hadley et al. reported on the assessment of operating room performance of four procedures in pediatric neurosurgery using an OSATS (Hadley, Lam et al. 2015). While there was no difference in scores given to junior and senior residents by faculty for novice procedures, there was a significant difference for expert procedures. While the study by Hadley et al. was underpowered to detect a difference (there were only 8 residents), and raters were not blinded, the authors determined that there was progression in the ability to perform more complicated tasks over time.

However, it is important to acknowledge that some junior residents were able to achieve a minimal level of competence at some of the more difficult tasks, potentially supporting the theory that there is a significant interaction between a learner’s level of interest and perceived value of a task, and their ability to acquire certain skills (Dunphy and Williamson 2004, Pusic, Boutis et al. 2015). It is also accepted that the journey towards competence is unique, dependent on the individual learner, with different skills developing at different rates depending on content, context, and past experience (Carraccio, Benson et al. 2008). While these factors were not specifically examined in this thesis, they cannot be disregarded as important factors, especially in the age of increasing subspecialisation. It seems logical that a resident who wishes to become a spine surgeon might not be particularly interested in becoming competent at arthroscopic labral repair, a procedure that they will not perform in their careers post graduation. For this reason, the issue of residents failing to achieve a minimal level of competency in the setting of the sports medicine rotation must be addressed.
There are two main approaches. Firstly, the establishment of critical competencies, such as EPAs, will be important for specialty training programs in the era of CBME – while the future spinal surgeon should not be expected to perform an arthroscopic labral repair, performing an ankle fracture fixation or hip fracture fixation is a common and important component of most surgeons’ trauma practice, requiring a minimal level of competence to be demonstrated by all trainees prior to graduation. The second (and complementary) approach is to use a milestones model as a developmental framework. This framework, as championed by ACGME, aims to pair competency assessment with milestones – in this manner the information gleaned from the studies in this thesis can be used to embrace different competencies for junior and senior resident, following a model that expects progression towards competence over time. Assessment tools are then used to ensure that these targets are being met.

The milestones approach is useful for both the assessment of clinical skills and the performance of technical procedures. For example, the results of the OSATS study have shown that it is reasonable to expect that junior residents are able to perform knee arthroscopy in a dry model to a minimal standard, with performance of rotator cuff surgery and arthroscopic knot tying a reasonable expectation of senior residents. The milestones approach is also a practical approach to the management of the myriad of orthopedic patient conditions, especially those typically managed by subspecialists with advanced training. To once again use the example of ACL patient care from the ACGME Orthopedic Milestones Project, the ultimate goal is a (senior) resident who can recognize concomitant injuries, interpret advanced imaging, and provide complex nonoperative patient management – a competent practitioner. Using this approach, the sports medicine rotation can ensure that residents demonstrate safe and competent management of patients presenting with injuries to their ACL, while recognizing that only a proportion of those residents who go on to subspecialty training should be expected to surgically manage these conditions and their resultant complications.

The results of the studies presented in this thesis provide evidence that experience and exposure are important components of the acquisition of clinical skill. Whether
the experience and exposure that is most important is related solely to sports medicine experience, or to overall experience from all rotations is unknown, but common sense dictates that both factors have an effect. The type and value of experiences obtained by residents over the years in other rotations is hard to quantify, in regards advances in clinical reasoning and psychomotor development, but it is likely an important effect, and one that underpinned the traditional time-based models of surgical training for decades.

There are two important conclusions from this thesis. Firstly, despite training within a CBME rotation, it cannot be expected that all junior residents will achieve a minimal level of competence. However, the studies in this thesis have demonstrate that there is a progression from novice to competent over time, allowing the progression of each resident to be mapped, ensuring that each is on the correct pathway to competence (Regehr 2010, Pusic, Boutis et al. 2015). In the future, on a group level, these types of curves might provide important information such as how many tasks repetitions are required to provide competence, or how much time is required (Pusic, Boutis et al. 2015). Unfortunately, teasing out the relative importance of each of these factors is beyond the scope of this thesis.

Secondly, it is crucial that resident seniority is not blindly associated with competence, a major failing of the traditional time-based programs. The information generated on the attainment of competence allows screening for resident learning problems, acknowledging those residents who are on the correct trajectory, and identifying those who are having difficulty. Other authors have recently published on the use of multiple assessment tools to evaluate surgical residents, using mapping of performance to successfully identify those residents who are outliers in various facets of training (Harrington, Miner et al. 2015). This thesis has served to highlight a major advantage of an outcomes-based program - using objective assessment tools to ensure that all residents are performing to a minimal level of competence, and to identify those who need remediation.
IDENTIFICATION OF THE RESIDENT IN DIFFICULTY

One of the proposed advantages of an outcomes-based program is the use of comprehensive, objective assessment tools to identify the resident in difficulty (Holmboe, Sherbino et al. 2010). While studying the acquisition of clinical skills and performance of technical procedures has given us an understanding of the progression of these skills that occurs over time, it has also allowed the identification of residents who are not performing to the level of their peer group, whether it be junior or senior residents.

This was seen most clearly in the clinical skills OSCE. Prior to using a test-centred standard setting method, over 60% of junior residents were not able to pass the OSCE – after the application of the modified Angoff to set a junior cut-score, only 1/21 junior residents failed the examination. This resident is an outlier with regards to their peer group – using objective criteria, they have been identified as underperforming in the area of clinical skills, compared to their peers, and would benefit from remediation.

The best form of remediation for the underperforming resident is unknown – currently remediation has taken the form of either extending the rotation by six or twelve weeks, or by the resident returning to repeat the rotation. The remediation usually occurs under the supervision of a different member of faculty, and is augmented by increased and regular use of WBAs, in addition to the repetition of the exit examinations. While WBA have many of the issues discussed in the literature review, their use in this setting serves to increase the level of faculty supervision, increasing the learner-centeredness of a CBME model.

While remediation is often focused on aspects of clinical and technical skills, it has been shown that the majority of residents in difficulty have issues that span more than one CanMEDS role (Zbieranowski, Takahashi et al. 2013, Guerrasio, Garrity et al. 2014). Yost et al. used an online assessment tool to characterize resident behavioral styles and motivators, and identified many factors that were associated with substandard performance on the American Board of Surgery In-Training Examination (ABSITE) (Yost, Gardner et al. 2015). While the CanMEDS OSCE was able to identify
residents with deficiencies in some of the CanMEDS roles, this type of assessment is not easily applicable to regular in-training assessment. At this time it is difficult to know how to address resident issues with roles other than Medical Expert, but remediation with other members of faculty, close faculty supervision, and regular use of multiple forms of assessment is an important first step.

There is evidence that, overall, resident remediation is generally successful in approximately 80% of cases (Zbieranowski, Takahashi et al. 2013, Guerrasio, Garrity et al. 2014). However, there may be a small percentage of residents who cannot be remediated, repeatedly fail different rotations, and may need to leave the training program to pursue alternative careers. While this scenario occurs rarely, there is evidence (as presented in the literature review) that current surgical training is failing some residents. Without objective evidence, provided over time by multiple observers using validated assessment tools, it is very difficult and even unfair to ask residents to consider alternative careers. Furthermore, such a move is inappropriate without training programs identifying these residents early, and offering a robust program of appropriate remediation. Regular assessment in the setting of CBE addresses both of these issues (Holmboe, Sherbino et al. 2010).

Interestingly, 1/6 of the sports medicine fellows were unable to pass the clinical skills OSCE. However, when the cut-scores generated by the modified Angoff and the BG / BLR methods were applied, this fellow passed the OSCE. Whether this approach is appropriate in the setting of a senior trainee / fellow is unknown – two experienced faculty deemed the fellow to be less than competent at 2/6 stations using a global rating scale, and only the total checklist score has allowed the fellow to pass. This raises two issues. Firstly, the argument surrounding checklists versus global ratings re-emerges, which has been greatly debated in the literature. The majority of authors believe that, overall, global ratings are important in the differentiation of performance, rewarding highly organized individuals, and identifying those who have not performed strongly, despite scoring well on a checklist (Norman, Van der Vleuten et al. 1991, Hodges, Regehr et al. 1999). Secondly, it raises the issue of ensuring that fellows
achieve a minimal level of competence, and the potentially important application of CBME to fellowship training. This is discussed further in Chapter 11, future directions.

THE RELEVANCE OF TIME

One of the proposed advantages of CBME is the potential to shorten the length of residency to the amount of time taken for a resident to complete various competencies – a proposal that has both supporters (ten Cate and Scheele 2007) and opponents (Brooks 2009). Reduced time in our residency program at this time has applied to a few residents, in whom frequent, summative assessment has allowed rapid advancement. Each of these residents had been deemed competent in each of the 21 rotations, and allowed to sit the orthopedic certification exam one year early. While the assessment tools in each of the rotations are not as robust as those used in the sports medicine rotation, it is important to note that after four years of training, each of these residents has passed the certifying examination that is administered by the RCPSC.

Shortening training must be viewed in the context of other changes in surgical training and practice. Most orthopedic surgeons do not have the general orthopedic practices that were common twenty or thirty years ago – the majority of young surgeons (>90%) are now are fellowship trained, and have subspecialty surgical practices (Horst, Choo et al. 2015). These changes have been driven by the recognition that many procedures undertaken by subspecialists with advanced training have reduced complication rates and better outcomes than those performed by generalists (Barrack, Krempec et al. 2013, Cahill, Pahys et al. 2014). Anecdotally, all of the graduating orthopedic residents over the last five years from the University of Toronto orthopedic training program have undertaken one, and usually two years of fellowship training. It might be argued that time spent training in their subspecialty of choice is more valuable than time spent in rotations that are not relevant to their future practice, and that today’s trainee is modifying their training to suit their own particular career aspirations. Another change at the University of Toronto Division of Orthopedics that has contributed to early graduation has been the removal of many rotations that were not critical to orthopedic training, the so-called off service rotations in medicine and
emergency for example. In this way, four years has now become the minimal training
time required to complete all rotations, if no remediation is required.

It is important to note that CBME can also lengthen training - a key feature of CBME,
whereby learners are allowed to progress at their own rate (Iobst, Sherbino et al.
2010). In a true CMBE model, some residents can achieve competency status earlier
than required, while others require increased educational time (Ebert and Fox 2014).
Certainly, in the case of several residents in the University of Toronto training
program, increased time in certain rotations has been mandated, due to perceived
substandard performance. In this instance, training reverts to the traditional model,
with the certification examination taken after the fifth year of training – as there is only
one exam sitting per year, there is no option to be certified between the fourth and fifth
years of training.

The discussion of whether CBME can shorten or lengthen training leads into one of
most important discussions surrounding the transition to CBME - the potential
logistical chaos that can be associated with a true focus on the learner. Allowing
residents to lengthen or shorten three-month rotations, or take four versus five years
to complete training makes rostering for clinics, operating rooms, and on call duties
extremely difficult. For this reason, operating a training program that is a hybrid model,
as proposed by the RCPSC (RCPSC 2014) makes the most sense – programs can
maintain a time-based program, with set periods in rotations, complemented by
objective assessments that ensure that resident milestones are being accomplished.
In this manner, training programs can continue to run smoothly, while maintaining the
major advantages of CBME training – ensuring that minimal levels of competency are
demonstrated, and identifying the resident in difficulty, allowing for appropriate
remediation.

HIGHLIGHTING DEFICIENCIES IN THE TRAINING PROGRAM

While the use of regular, objective assessment in this thesis has provided
opportunities for both formative and summative feedback, it has also identified areas
for improvement in training programs. One example that emerged from this thesis was
the non-uniformity of training within the sports medicine rotation, as a result of faculty having different surgical practices. Feedback following their OSATS and OSCE assessments revealed that some residents felt that there had been insufficient exposure to knee and/or shoulder clinical scenarios and technical procedures. As a result of this feedback, two changes were made to the sports medicine rotations. Firstly, the clinical practices of all surgeons in sports medicine division were reviewed, allowing adjustment of some resident rotations to improve exposure to the critical competencies. Secondly, the number of training sessions in the skills laboratory was increased, in order to reinforce these critical competencies and to provide adequate opportunities for residents to improve their performance of technical procedures.

Another example of the importance of using assessment tools was seen with the EPA study. In this study, many junior residents and some senior residents were found to be less than competent in technical procedures such as fixation of an ankle or hip fractures, identified by faculty as critical competencies, despite all junior residents having been determined to be competent at basic trauma and hip fracture modules. This study demonstrates that blueprinting must be used to align rotations with critical EPAs, making specific procedures a focus of rotations, and using objective assessment methods ensure competence before residents leave a rotation.

THE SUPERIORITY OF CBME OVER TRADITIONAL TIME-BASED TRAINING

There is no doubting that regular assessment, allowing residents’ progression to be charted, and deficiencies identified, is advantageous to both resident and program. Outcomes-based postgraduate training also clearly increases faculty engagement, with regular teaching and assessment linked to carefully planned curricula. This results in regular feedback, providing insight into a resident’s skills and limitations, and allowing self-reflection (Bhatti and Cummings 2007). Despite these advantages, the evidence that CBME improves postgraduate training in comparison to time-based training is limited.

The results of this thesis are not able to provide evidence of superiority of CBME. Doing so would require a direct comparison of two groups – one group of residents
who are in the CBME program, and another who remain in a time-based program (and who are somehow not exposed to the teaching and assessment programs of the CBME). Such a study will not be possible at the University of Toronto – the CBME program is widely accepted by both the residents and faculty, and it might even be deemed unethical to prevent residents from being exposed to these important educational opportunities. Furthermore, a study that effectively compares time-based and outcomes-based training programs would require an outcome measure that is able to differentiate between senior trainees. At this time, the only outcome measure available is the orthopedic certifying exam, which the vast majority of residents pass.

There is some evidence from this thesis demonstrating that, at a minimum, the CBME training program at the University of Toronto is training residents appropriately. The clinical skills OSCE showed that senior residents were achieving a minimal level of competence in clinical skills, while the EPA study demonstrated that the majority of senior residents were able to perform the entrustable professional activities competently, in the simulated setting. Furthermore, the CanMEDS OSCE showed that senior residents were able to manage difficult clinical scenarios in an appropriate manner. While these studies do not provide evidence that implementing a CBME program is responsible for these effects, the use of regular assessment to demonstrate resident competency has provided a level of accountability that was missing from time-based programs. By implementing assessment programs that ensure residents are on the correct trajectory in the era of subspecialisation, and that ensure residents are achieving a minimal level of competency in critical competencies, CBME enforces a minimum standard that both governing bodies and the public expect.

9.3 LIMITATIONS

EVIDENCE OF VALIDITY

A common fault of many studies published in the field of surgical simulation is that they fail to use a modern framework of validity (Ghaderi, Manji et al. 2015). The results of four of the five studies presented in this thesis (excluding Chapter 5
discussing standard setting) show evidence of validity to varying degrees – generally speaking, the content and response process was acceptable for each of these studies. However, there was some room for improvement in aspects of each study.

*Internal structure*

Generally, the internal consistency of each of the studies in this thesis was very high, with two possible explanations. Firstly, in-training assessment by its very nature focuses on a narrow construct, which is known to increase reliability. Another explanation for the consistently high reliability may also partly be the effect of known-group comparisons (expert-novice comparisons), whereby including groups with known differences can inflate the reliability coefficient (Cook 2015). Certainly, the clinical skills OSCE and OSATS studies included fellows and staff, and compared these groups to novices. However, it might have been expected that the junior and senior residents groups would have been more homogenous, having been exposed to similar training – it might be argued that these studies compared two groups with similar training but unknown abilities, and have used the assessment tool to determine the ability of each individual (Cook 2015).

The OSATS and EPA studies reported on inter-rater reliability, which was high in both papers. Reasons for this high inter-rater reliability might include using experienced examiners, who had participated previously in OSCE and OSATS studies, as well as the use of carefully constructed checklists and previously validated global rating scales. However, no measure of inter-rater reliability was available in the clinical skills and CanMEDS OSCE studies, as neither second raters nor videotaping was used. While there is some evidence in orthopedic OSCEs that inter-rater reliability is acceptable (Mohtadi, Harasym et al. 1995, Vivekananda-Schmidt, Lewis et al. 2007, Chen, Lee et al. 2013), it cannot be assumed that the inter-rater reliability was high in these studies. With regards the CanMEDS OSCE, it has previously been suggested that the reliability of interpersonal skills assessment is reduced in comparison to clinical skills stations (Brannick, Erol-Korkmaz et al. 2011).
Another method of approaching issues of reliability is to use G theory. Throughout this thesis, classical test theory (in the form of Cronbach’s alpha) was used to calculate the internal consistency, a measure of how closely related a set of items are as a group. As discussed in the literature review, G theory can also be used to calculate reliability, and to take into account the variance component of all variables of interest.

The use of G theory to consider sources of error can be very useful in psychometric studies, whereby the effect of multiple factors such as raters, items, cases and subjects can be calculated (Bilgic, Watanabe et al. 2015). Generalizability theory was used in both the clinical skills OSCE and the OSATS, identifying that the major source of variance was the training level of the participant, rather than a result of raters or cases. However, G theory was not used in the CanMEDS study and the EPA, where again it would have been useful to identify potential sources of error in the rating of candidates. In the EPA study, for 6/9 stations the inter-observer reliability was found to be high, suggesting that the effect of rater error was minimal. However, no inter-observer reliability was available for the CanMEDS study – certainly G theory would have been useful to estimate the effect of any error of raters, compared to stations or items.

Generalizability theory was used in the OSATS study to calculate interstation reliability – the result was 0.78, which was less than the calculated alpha coefficient of 0.9. This finding has previously been reported - in a systematic review of OSCE reliability, the mean of generalizability coefficients was lower than the mean of the alpha coefficients (Brannick, Erol-Korkmaz et al. 2011). The authors of this systematic review postulated that the difference was a function of generalizability coefficients considering absolute differences to be meaningful, while alpha coefficients consider relative differences to be meaningful. Certainly, it would have been interesting to identify if there was a similar difference in the calculated reliability of the CanMEDS and clinical skills OSCE, and the EPA study.
**Relation to other variables**

Correlation with other scores from other instruments is an important source of validity evidence, and in this thesis was sought in various ways. In the CanMEDS OSCE, a correlation between ITER scores / program director rankings and performance was both sought and identified, while in the OSATS study a correlation was identified between performance of technical procedures, and the OSCE scores for each resident. These findings have been discussed above.

The clinical skills OSCE presented in Chapter 4 did not seek to identify a correlation between performance and other variables. In the study by Dwyer et al. (Appendix C), a good correlation was seen between the Orthopedic In-Training Examination (an MCQ test that all residents take) and performance on the sports medicine OSCE (Dwyer, Theodoropoulos et al. 2013). This comparison was not repeated, as there was no reason to believe that there would be a different result. However, as the assessment program for CMBE develops, it will be important to include other measures of clinical skills; options include open-ended questions such as short answer questions, and close-ended questions such as MCQ. While short answer questions have some merit, the burden of marking them, in association with evidence of low inter-rater reliability means that they should, in general, be avoided (Hift 2014). It is generally accepted that MCQ have a higher reliability than short answer questions, and are an efficient form of assessment, with evidence to support their use in both undergraduate (Hift 2014), and postgraduate examinations (Yen, Athwal et al. 2014, Kelly, London et al. 2015, Shetty, Grajo et al. 2015). For this reason, the assessment program in the sports medicine rotation now incorporates MCQ exams as part of the end of rotation assessment of clinical skills.

The ACLR study by Dwyer et al., which studied the use of a dry model to assess performance of a technical procedure, sought and identified a correlation between previous exposure to knee arthroscopy and ACL reconstruction, and performance (Dwyer, Slade Shantz et al. 2015) (Appendix D). A limitation of both the OSATS and the EPA studies was that neither provided details on the operative exposure of each resident, either study specific procedures or overall experience. The first reason for
this omission was that the information used in the ACLR study was based upon resident recall, which likely suffered from confounders such as recall bias. Secondly, while there is some use of logbooks in the residency program at the University of Toronto, these are not universally used, and their accuracy is unknown (Achuthan, Grover et al. 2006). While it seems logical that increased exposure to ACL reconstruction improves performance on a dry model, teasing out the overall operative experience of each resident might have helped to explain whether it was exposure to sports medicine procedures, or overall operative experience from all rotations that contributed to the superior performance of senior residents compared to juniors.

Importantly, the OSATS study and the EPA study did not provide any correlation between performance of the procedures on the dry models, and operative experience. While there are many studies (presented in the literature review) demonstrating that bench model training improves operative performance in orthopedics and other surgical specialties, the evidence that simulation-based assessment correlates with actual surgical performance is more limited (Korndorffer, Kasten et al. 2010), although the findings of the few studies available consistently identify a good correlation (Datta, Bann et al. 2004, Beard, Jolly et al. 2005, McCluney, Vassiliou et al. 2007, Ghaderi, Vaillancourt et al. 2011). The reasons for this deficiency in the literature is likely because of the reasons discussed in the literature review, including difficulty standardizing operations, time restraints, issues of patient safety, as well as the logistical difficulties faced in obtaining resident ratings from faculty who are not familiar with the trainee. However, this will be a critical aspect of our next phase of study – finding a way to measure intra-operative performance of critical orthopedic procedures in a reliable way.

Another limitation was the limited correlation of performance in the sports medicine OSCE and OSATS with performance in other rotations. While the EPA study demonstrated that senior residents outperformed junior resident on tasks critical to basic trauma, hip fracture and basic arthroplasty rotations, it cannot be assumed that senior residents will outperform junior residents in other rotations, however likely it may seem. This will be another important area of future research - using the
assessment tools studied in this thesis, and applying them to other rotations, it may be that there are clinical skills and technical procedures that can be mastered by junior residents.

Consequences

It is important that the results of scores from assessment tools are used to make decisions, or to impact teaching and learning (Cook, Beckman et al. 2010, Cook 2015). The studies presented in this thesis were usually the first of their kind in the setting of CMBE in our orthopedic training program – without evidence of validity it was, in the beginning, felt inappropriate to use these results to determine whether residents should repeat rotations. The results of the clinical skills OSCE demonstrate the importance of analyzing results prior to making relatively important decisions - without the modified Angoff standard setting method, an unacceptable number of junior residents would have failed. The ramifications of the OSATS study would have been even more significant, with many junior and senior residents unable to perform the technical procedures competently.

Despite this, consequential validity is important - the scores that are attained from assessment tools must be used in a meaningful way. In some ways, this has occurred. The impact on teaching and the training program has been profound – a milestones approach has been implemented in the clinical skills OSCE, with a different standard for junior and senior residents expected. The technical procedures that junior residents are expected to be competent at have been modified, as have those for the senior residents. The CanMEDS OSCE is now being used on an annual basis; residents not found to be performing competently are now expected to undergo remediation.

Finally, the EPA study demonstrated that many junior and some senior residents were not able to achieve a minimal level of competency in three critical competencies of orthopedics. This finding must have an educational impact - if lists of EPAs are developed for specialties, objective assessment for each of these must be used to demonstrate a minimal level of competence in all residents prior to graduation.
Furthermore, the rotations during which residents are trained in these skills must focus their attention upon performance of these critical tasks, and ensure that teaching and assessment vigorously matches the curriculum.

**POWER AND BIAS**

Power analyses were performed in the clinical skills OSCE study (which was underpowered to detect any significant difference between the fellows and senior residents), and in the EPA study (which was underpowered to detect differences between the PGY-1 and the PGY-4 groups). A power analysis was not performed in either the CanMEDS OSCE or the OSATS study – while both of these studies identified important differences between groups, the CanMEDS OSCE was unable to identify a difference between the PGY-0 and the PGY-3 residents, despite a trend for increased scores in the PGY-3 group. It seems likely that, due to limited numbers, this latter study was underpowered to do so.

Lack of statistical power to detect important differences is commonplace in simulation studies, usually due to the limited number of subjects (students, trainees, residents) available for study. While it is often difficult to increase subject numbers in the behavioral sciences without conducting multi-centred trials, it is important to recognize when a study is underpowered. In this way, a Type II error can be avoided – i.e. believing that there is no difference between groups when in fact there might be. However, the studies in this thesis did detect important differences between groups, despite the relatively limited numbers.

A major potential issue in the clinical skills OSCE and the CanMEDS study was one of potential bias, whereby faculty awareness of year of training, or familiarity with residents may have influenced examiner ratings. The OSATS and EPA studies used videotaping to measure inter-rater reliability, which was shown to be high in both studies – as a result it seems likely that the influence of bias was relatively low. However, inter-rater reliability was not available for either the clinical skills or the CanMEDS OSCE, as only one examiner was used in each station. In these studies, some of the examiners were aware of the year of training of some of the residents,
raising the potential for leniency and stringency as a source of error. In this way, some examiners might have been more lenient on senior residents, more stringent on junior residents, or vice versa.

Evidence for the effect of bias was sought in the clinical skills OSCE, with subgroup analysis demonstrating an effect of familiarity of examiners with residents in only one of six stations. Whether this was a true finding or random chance is unknown. However, it is not certain that such bias, if it existed, would serve to generate the finding that senior residents outperform junior residents, as it seems likely that familiarity would affect senior and junior resident scores in a similar fashion. It is also important to note that that the majority of examiners in the clinical skills and CanMEDS OSCEs were not aware of the research question of this thesis and had no vested interest in finding a difference between junior and senior residents. It could actually be argued that, because many of the faculty were educators heavily involved in the development and implementation of the CBME program, it is just as likely that these examiners would be reluctant to fail significant numbers of junior residents.

Once again, it is important to acknowledge that true objectivity is very difficult to obtain in most assessments, as the assessment of competency always requires the exercise of judgment by another competent, proficient or expert practitioner (Brooks 2009). An attempt to minimize these subjective effects is made via processes of standardization, by reaching consent regarding what is being assessed, and the standards expected (Ginsburg, McIlroy et al. 2010). While high-stakes examinations avoid pairing examiners with familiar candidates, this is often impossible in low- or medium-stakes examinations (such as the studies in this thesis) due to cost, examiner availability and scheduling difficulties (Stroud, Herold et al. 2011). Our experienced examiners were asked to disregard the PGY level of residents, and it would seem that they were reasonably effective at doing so. Future options to minimise the effect of bias would be to use faculty from a separate institution, an option not typically feasible in the setting of regular assessment, or to provide evidence of inter-rater reliability by videotaping all encounters, also not particularly feasible.
THE EFFECT OF THREE- VERSUS SIX-MONTH ROTATIONS

While a major finding of this thesis was that there were important differences between junior and senior residents after a sports medicine rotation, it must be acknowledged deciphering the main cause of this finding is difficult. Potential reasons include the experience, exposure and skills gained through years of training in all rotations, or the experience and skills gained in sports medicine rotations, or a combination of the two.

The sports medicine training at the University of Toronto is comprised of two three-month rotations – the first as a junior resident (PGY-1,2&3) and once as a senior resident (PGY-4&5). None of the senior residents had undertaken either the OSCE or the OSATS prior to participating as part of this study. The exam was targeted at a senior level or knowledge, or at the knowledge level of a practicing, general orthopaedic surgeon – both junior and senior residents were aware that this would be the case. A critical question is whether junior residents, if allowed to undertake a six-month rotation, would be able to achieve a minimal level of competency, or a level of performance on a par with that of senior residents? Is this a dose response finding? As identified in some of the technical procedures studies in the literature, as well as the ACLR study included in Appendix D, performance has been associated with both the number of procedures performed, as well as time in training (Pugh, Hamstra et al. 2015). The two are not mutually exclusive.

Determining whether or not junior residents would achieve a minimal level of competency if they were to undertake the sports medicine rotation for six rather than three months would require a major change to the curriculum map. In order to answer this question, one group of residents would need to undertake two three-month rotations, while a second group of residents would undertake a single six-month rotation, and a comparison made between the two groups. However, it may be simpler to conclude that residents cannot achieve a minimal level of competence after three month rotations in subspecialties that surgeons devote their entire careers to, and concentrate on developing lists of critical competencies and milestones that residents should be expected to perform and achieve.
The results of this thesis tell us that the senior resident who returns at the end of training for a second rotation is more likely to be competent than the junior resident undertaking a first rotation. Whether the difference in performance between junior and senior residents is the result of increased time in training, increased training in sports medicine, increased exposure to sports medicine procedures, increased exposure to all orthopedic procedures, or as a result of senior residents being more motivated is unknown. It also can be argued that it does not matter. It is more important that educators are using objective, validated assessment tools to ensure residents achieve a minimal level of competence in managing clinical scenarios, and performing those procedures deemed critical to training.

THE USE OF SIMULATION V WORKPLACE-BASED ASSESSMENT

Finally, a criticism of this thesis could centre on the use of simulation-based assessment, as opposed to WBA. It would be hard to deny that the performance of clinical skills and technical procedures by residents should also be assessed in the workplace. While it is unlikely that a resident unable to manage a virtual patient or perform a simulated procedure would be able to do so in the workplace, competent performance in a simulated setting does not ensure that a resident can perform competently in the clinic or the operating room. For this reason, assessment in the workplace must be an important component of CBME.

The mini-CEX is commonly used to assess performance of clinical skills, and well accepted in the setting of postgraduate medical training (Wiles, Dawson et al. 2007, Liao, Pu et al. 2013). However, issues with the mini-CEX, like all assessment tools, have been raised, including rater error, insufficient content expertise of raters in the skills being assessed, and variable frames of reference (Kogan, Conforti et al. 2011, Berendonk, Stalmeijer et al. 2013, Yeates, O'Neill et al. 2013). Another critical issue that has been highlighted in the use of WBA is one of feasibility – in a cardiology training program, residents were required to provide four documented mini-CEX (Alves de Lima, Barrero et al. 2007). Over an 18 month period, only 15% of residents able to do so - the results of this study likely speak to issues of faculty engagement and
assessment overload, as well as the administrative issues that revolve around ensuring WBA are performed, documented, and the results used in a meaningful way.

The use of OSATS, Direct Observation of Procedural Skills (DOPS), and other methods to assess the performance of residents in the operating room continue to be reported in the literature (Awad, Hayden et al. 2014, Hopmans, den Hoed et al. 2014, Ahmidi, Poddar et al. 2015, Hadley, Lam et al. 2015, Husslein, Shirreff et al. 2015, Kramp, van Det et al. 2015, Nandigam, Soh et al. 2015). While operative performance is a defining characteristic of surgeons (Williams, Verhulst et al. 2012), widespread implementation of WBA as it pertains to the operating room is slow, even at the University of Toronto. The reason for this is relatively simple. As Anastakis et al. elegantly summarized, “cases vary widely with respect to difficulty and complications, and faculty vary widely on the amount of independence they are willing to allow a trainee in their operating room” (Anastakis, Regehr et al. 1999).

Other issues have been identified with the use of assessment in the operating room. Using a generalizability analyses, Williams et al. identified that rating between two and three operative procedures every month was required to obtain a reliability of 0.8 (Williams, Verhulst et al. 2012). While these numbers might not seem overly arduous, two or three operative assessments must be multiplied by the number of residents, as well as the number of procedures deemed critical to each specialty. The solution will be to use a combination of both simulation as well as workplace assessment. If the advantages and disadvantages of simulation-based assessment and WBA are both acknowledged and minimized, these methods can be used together to provide an overall picture of a resident’s progression towards competence, over years in training rather than months in a rotation.
Chapter 10

Conclusion

A combination of assessment tools assessing clinical skills, the application of the intrinsic CanMEDS roles, and the performance of technical procedures demonstrated sufficient evidence of validity for use in the setting of in-training assessment. These assessments consistently demonstrated that many junior residents, in comparison to senior residents, were not able to achieve a minimal level of competence. Implications of these findings affect the fundamental concept of achieving competence before progressing to the next rotation in CBME, and lend credibility to hybrid versions of CBME whereby both time-based and outcomes-based models are combined. The findings of this thesis emphasize the importance of a milestones approach to postgraduate training, as well as the importance of ensuring that residents are able to perform the critical tasks of each specialty to a competent level, prior to graduation.
Chapter 11
Future Directions

WHAT ASSESSMENT TOOLS SHOULD BE USED IN THE FUTURE?

The clinical skills OSCE worked well as an in-training assessment tool, and was shown to be useful in differentiating residents able to perform competently from those who were not. However, an OSCE model will not be suitable for use in all 21 rotations, especially those that have only one resident at a time. For these rotations, it is likely that the continued use of an oral examination will be most suitable, although ideally the examiner would not be the faculty supervisor for that rotation. However, the evidence from this thesis supports the continued use of an OSCE for important rotations, such as trauma and arthroplasty that have large numbers of residents at any given time. It also seems likely that an OSCE should be used in conjunction with a standard setting method, used to set different cut scores for junior residents should this prove necessary. The assessment of clinical skills in the majority of rotations will also benefit from the addition of assessment tools such as an MCQ, and mini-CEX. Once question banks are created, MCQ exams generally have high reliability, are easy to both mark and analyze, and serve to broaden the assessment profile of residents. While there are issues with the mini-CEX, this should be added to selected rotations in order to provide some WBA to complement the other assessment tools.

The assessment of the intrinsic CanMEDS roles continues to be problematic, despite the success of the CanMEDS OSCE, now used once a year. The frequency of this cannot be increased, due to cost. One option would be the increased use of 360-degree evaluations – at this time these are not being used to their full potential, likely due to logistical issues. Another is to include CanMEDS stations in end-of-rotation OSCEs, which is currently happening in the arthroplasty OSCE. While having a single station is not likely to be a reliable method of assessing the intrinsic roles, adapting these stations to be predominantly formative rather than summative has been working well so far. It must be pointed out that the current curriculum map (Appendix B) for
the Division of Orthopedics has different roles for junior and senior residents for each rotation – it is likely that this is unnecessary, and complicates the process of both teaching and assessing these roles. A single role that is the focus for each rotation would likely suffice.

As previously mentioned, the assessment of technical procedures continues to pose difficulties. Despite the success of the OSATS study, issues such as cost and faculty time will likely limit the use of an OSATS model to only a few of the 21 rotations. Clearly, there needs to be an increase in the assessment of operating room performance in many of the CBME rotations. One method than can provide guidance in would be to formally decide on a list of critical EPAs in orthopedics. By determining the most important 20 or 30 technical procedures, these can be mapped to the curriculum, and assessment tools used to ensure a minimal level of competence. Generating a list of EPAs will also change the focus of assessment away from complex procedures that many residents will not perform after graduation, to those that are a critical component of orthopedic practice. In this way, resources such as money and time can be used optimally in training.

Another major issue is which faculty members should perform the workplace-based assessment, especially as they pertain to technical procedures. While it is not ideal to have the faculty supervisor perform the operating room assessment, logistically it is difficult to arrange alternative faculty members to rate residents. Doing so is more feasible in the setting of simulation-based assessments such as the OSATS or the EPA study, in which the date and time of each assessment are easily controlled. One possibility would be for residents to change supervisors in the last two weeks of their rotation, however this solution likely has limited feasibility and acceptability to those involved, and increases the complexity of rostering. For these reasons, it seems that supervising faculty, for the most part, will continue to perform resident assessment in the operating room at the end of rotations – careful selection of procedures, the use of validated assessment forms, and appropriate faculty training will help to ensure that subjectivity is minimized.
RECOMMENDATIONS IF WE ACCEPT JUNIOR RESIDENTS CANNOT BECOME COMPETENT AFTER A SINGLE ROTATION IN CBME

The finding that many junior residents cannot achieve a minimal level of competence is limited to the sports medicine rotation, although the EPA study demonstrated similar findings in technical procedures that comprise important components of the basic fracture, hip fracture, and arthroplasty rotations. Further research is required to determine if these findings are encountered after other rotations in the orthopedic training program. However, in light of the evidence that has been presented, it seems extremely unlikely junior residents will be able to achieve a minimal level of competence after short rotations in modules as complex as spinal surgery and pediatric orthopedics.

Two clear options have emerged from this thesis, and both should be applied. The first is to determine a list of critical EPA’s for a specialty, focus teaching and training on these competencies during the relevant rotations, and apply objective assessment tools in order to identify deficiencies and allow remediation. For all other competencies, I believe that milestones should be used as a developmental framework, and assessment tools used to ensure that residents are progressing through these predetermined milestones.

The results of the clinical skills OSCE and the OSATS demonstrated the importance of using milestones to chart the development of residents’ clinical abilities over time. Using the Dreyfus model of skill acquisition, it was clear that the majority of residents progressed in an upward trajectory over time, in a manner similar to that outlined by the ACGME Orthopedic Milestones Project. The continued use of assessment tools in a CBME program should be to ensure that each resident is on a trajectory consistent with appropriate milestones, moving away from the overly simplistic concept of determining if residents are competent or not competent after a rotation.

The motivation for this thesis was to examine, and try to dispel, the concept that all residents can become competent at various orthopedic subspecialties after short training rotations. It seemed inappropriate to describe any resident as being
competent, without valid and objective assessment tools. Furthermore, to describe a junior resident as competent after a subspecialty rotation, especially those subspecialties that require fellowship training, inspired a degree of incredulity in many of the faculty not directly involved in the implementation of the CBME program. For this reason, in order to encourage widespread acceptance of the CBME model, which has many advantages, redefining the concept of competency is important. I believe that it is much more appropriate to describe a resident as being competent to perform an EPA, or as having achieved a milestone appropriate for their level of training. In this way, a CBME program can be used to potentially improve training, improve accountability, and most importantly, identify the resident in difficulty.

HOW MIGHT WE DEMONSTRATE THAT CBME IS SUPERIOR TO TRADITIONAL TRAINING?

Proving that CBME is superior to a traditional time-based training will be difficult to do, without restructuring a training program to have both training methods run concurrently at a single institution, with residents randomly assigned to participate in one or the other. Furthermore, as previously mentioned, sufficiently detailed outcomes measures to demonstrate any potential difference are lacking. Certainly, the assessment tools discussed in this thesis could be used throughout each of the training programs, however if regular and objective assessments were to be applied to both programs, what would differentiate the two? Qualitative research is one option – by studying residents who have and have not been part of the CBME program at the University of Toronto, as well as faculty members involved in the implementation of CBME, important themes may emerge. It is common for residents in the sports rotation to provide feedback that includes appreciation of sustained and focused teaching, appreciation of procedural training on dry models, as well as appreciation of the OSCE assessments, as most residents accept that assessment drives learning, and value exposure to and practice of exam techniques.

There may be little need to demonstrate the superiority of CBME through carefully constructed, costly and resource heavy trials. The CBME program at the University of Toronto has altered the training landscape – the curriculum has been carefully
restructured, ensuring exposure of all residents to the critical competencies of our specialty. Teaching and training is focused, resulting in increased faculty involvement, and increased attention to the performance of technical procedures using simulation. Objective assessment is being used on a regular basis to provide feedback for those residents identified as progressing appropriately, and allowing the identification of residents in difficulty, enabling appropriate remediation. The increased focus on the learner is clearly beneficial to the resident, the program, and the public.

THE END OF CERTIFICATION EXAMINATIONS

This thesis has demonstrated that a combination of assessment tools can be used to determine a minimal level of competency in residents, and chart resident progression over time. As these assessment tools, which have been carefully selected to maximize feasibility, become common in the CBME program at the University of Toronto Division of Orthopedics, the wealth of information available may change the education landscape. With each resident’s knowledge and technical skills being consistently and objectively assessed after each of 21 rotations, there may come a time where the orthopedic certification examination (and other residency certification examinations) become both superfluous and unnecessary. Not only will residents have been tested on their knowledge and decision-making after multiple rotations, and their application and understanding of the intrinsic CanMEDS Roles assessed, but residents will have also been required to demonstrate their ability to perform technical procedures, currently beyond the scope of the certification examinations (and not likely to change).

The cost savings of such a shift are likely to be substantial – each year an average of between 60 and 70 final year orthopedic residents undertake the certification examination across Canada. Up to 40 faculty are required to act as examiners, while exam creation takes up substantial amounts of faculty time, with an associated and substantial financial cost. Should the expected transition from traditional time-based training programs to CBME occur in all residency training programs across Canada, in association with the use of objective, validated assessment tools frequently throughout
residency, certification examination will no longer be required. In this way, the burden of assessment will move from the RCPSC to residency programs.

FUTURE SCOPE FOR CBME

To date, the focus of CBME has been on residency training. However, the evidence from this thesis suggests that many residents cannot achieve a minimal level of competence – this was most evident in the sports medicine technical procedures, which many senior residents had difficulty performing even in the simulated setting. There are many reasons for this. The first is that orthopedics is an extremely diverse specialty – using the sports medicine subspecialty as an example, not only do surgeons in this field perform a limited number of procedures, and often manage a limited number of conditions, but many are super specialized, performing only knee, shoulder, hip or elbow surgery due to the complexity of both procedures and conditions. This change in practice to subspecialisation and super-specialization has been driven by significant changes in technology and procedures over the last few decades. Furthermore, both surgeons and patients understand that, for the most part, it is impossible to be expert at hundreds of procedures - fellowship training and subspecialisation is required in order to develop and maintain proficiency and expertise.

For these reasons, the vast majority of graduates from orthopedic residency programs will pursue fellowship training and have subspecialty careers. However fellows function in a traditional time-based training program – competence is synonymous with having spent a year in an apprenticeship model. At the end of this year, usually without any assessment at all, a certificate is awarded enabling a surgeon is able to perform complex procedures. It is assumed that the fellow can perform to a competent or proficient level - whether that confidence is well based is unknown. For this reason, the integral concept of an outcomes-based program, whereby a minimal level of competence must be demonstrated prior to certification, must be applied to fellowship training
This is the logical progression of outcomes-based medical education – to focus assessment on surgeons, asking them to demonstrate a minimal level of competence in those clinical skills and technical procedures that they plan to perform. This focus on assessment and demonstration of technical competence has already begun in some subspecialty training programs, such as colorectal surgery (de Montbrun, Roberts et al. 2013). This is not to disregard CMBE’s important place in residency training, but rather to recognize that many training programs aim to give trainees broad experience in all aspects of a specialty, with an expectation that subspecialty training will follow. As with CBME in residency, the development of critical competencies for each orthopedic subspecialty will help guide curriculum development, and the optimal assessment tools. In this way, all surgeons wishing to pursue a particular practice would be required to demonstrate a minimal level of competency, providing a new and much needed level of accountability.
Appendices

Appendix A. Curriculum map for Competency-based Curriculum at the University of Toronto Division of Orthopedics

<table>
<thead>
<tr>
<th>TABLE II Curriculum Map for Competency-Based Curriculum Detailing CanMEDS Competencies That Are Formally Evaluated in Each Module</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase I</strong></td>
</tr>
<tr>
<td>Module 1</td>
</tr>
<tr>
<td>Module 2</td>
</tr>
<tr>
<td>Module 3</td>
</tr>
<tr>
<td>Module 4*</td>
</tr>
<tr>
<td>Module 5</td>
</tr>
<tr>
<td>Module 6</td>
</tr>
<tr>
<td>Module 7*</td>
</tr>
<tr>
<td>Module 8*</td>
</tr>
<tr>
<td>Module 9*</td>
</tr>
<tr>
<td><strong>Phase II</strong></td>
</tr>
<tr>
<td>Module 10</td>
</tr>
<tr>
<td>Module 11</td>
</tr>
<tr>
<td>Module 12</td>
</tr>
<tr>
<td>Module 13*</td>
</tr>
<tr>
<td>Module 14</td>
</tr>
<tr>
<td>Module 15</td>
</tr>
<tr>
<td><strong>Phase III</strong></td>
</tr>
<tr>
<td>Module 16</td>
</tr>
<tr>
<td>Module 17</td>
</tr>
<tr>
<td>Module 18</td>
</tr>
<tr>
<td>Module 19</td>
</tr>
<tr>
<td>Module 20</td>
</tr>
<tr>
<td>Module 21</td>
</tr>
</tbody>
</table>

*The module was covered by external rotation and/or a university course.
APPENDIX B: Curriculum overview map, with the intrinsic CanMEDS Roles mapped to each rotation

<table>
<thead>
<tr>
<th>PHASE 1</th>
<th>Module 1</th>
<th>Module 2</th>
<th>Module 3</th>
<th>Module 4</th>
<th>Module 5</th>
<th>Module 6</th>
<th>Module 7</th>
<th>Module 8</th>
<th>Module 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intro</td>
<td>Hip and Basic Fractures</td>
<td>Emergency Surgery</td>
<td>Medical Comorbidities in the Surgical Patient</td>
<td>Arthroscopy</td>
<td>Arthroplasty</td>
<td>ICU</td>
<td>Core Training in Surgery</td>
<td>CanMEDS core comp'</td>
<td></td>
</tr>
<tr>
<td>Prof</td>
<td>Sch</td>
<td>HA</td>
<td>C</td>
<td>Com</td>
<td>Prof</td>
<td>Man</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PHASE 2</th>
<th>Module 10</th>
<th>Module 11</th>
<th>Module 12</th>
<th>Module 13</th>
<th>Module 14</th>
<th>Module 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paediatric fractures</td>
<td>Spine</td>
<td>Foot and ankle</td>
<td>Basic Science</td>
<td>Hand and upper extremity</td>
<td>MSK medicine</td>
<td></td>
</tr>
<tr>
<td>HA</td>
<td>Com</td>
<td>C</td>
<td>Man</td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sch</td>
<td>Prof</td>
<td>HA</td>
<td>Com</td>
<td>Sch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PHASE 3</th>
<th>Module 16</th>
<th>Module 17</th>
<th>Module 18</th>
<th>Module 19</th>
<th>Module 20</th>
<th>Module 21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oncology</td>
<td>Complex trauma</td>
<td>Complex arthroplasty/joint reconstruction</td>
<td>Paediatric orthopaedics</td>
<td>Sports</td>
<td>Research/Flexible Experiences</td>
<td></td>
</tr>
<tr>
<td>Prof</td>
<td>Man</td>
<td>Man</td>
<td>Prof</td>
<td>Sch</td>
<td>Sch</td>
<td></td>
</tr>
<tr>
<td>Com</td>
<td>Com</td>
<td>HA</td>
<td>C</td>
<td>C</td>
<td>Sch</td>
<td></td>
</tr>
</tbody>
</table>

Module covered by external rotation &/or university course
APPENDIX C: Assessing competence of orthopaedic residents: The reliability and validity of an Objective Structured Clinical Examination (OSCE) after a sports medicine rotation.

Dwyer T, Theodoropoulos JS, Herold J, Henry P, Wasserstein D, Murnaghan ML, Wadey V, Hodges B, Semple J, Ogilvie-Harris D.


ABSTRACT

Background

The development of competency-based curricula in Canadian orthopedic resident training requires an objective means of assessing in-training competence after a module. We hypothesized that an orthopaedic sports medicine Objective Structured Clinical Examination (OSCE) would demonstrate sufficient reliability and validity to be used as an assessment of resident in-training competence.

Methods

An OSCE made up of six ten-minute stations was written for orthopedic residents, assessing resident competencies in the management of common sports medicine conditions. Each station covered domains of history taking, examination, imaging, clinical decision-making, consent and surgical technique. A combination of binary checklists and overall global ratings were used. Results were analyzed for overall reliability and construct validity.

Results

A total of 43 residents sat the OSCE. Cronbach's (Inter-station) Alpha was 0.91 for the six stations, with each station having acceptable Corrected Item-Total Correlation Coefficients. A significant effect of year of training was seen on both total checklist scores and global ratings, as well as on all individual domains of performance except
history. Final year trainees performed particularly well in clinical decision-making and knowledge of surgical techniques. Overall, the number of previous orthopaedic sports medicine rotations that a resident had undertaken correlated with significant increases in checklist scores, while recent exposure to competency-based sports medicine rotation did not.

**Conclusion**

An orthopedic sports medicine OSCE has shown sufficient validity and reliability to be used as an in-training assessment tool. Issues of standard setting must now be addressed in order to determine competence.

**INTRODUCTION**

The traditional method of orthopedic training in Canada utilizes a time-based system, combined with a formal exit examination. Increasing interest in competency-based curriculum has raised the issue of how to test in-training competence. Currently, the most commonly used method is the In-Training Evaluation Report (ITER), which is known to be relatively subjective in nature; the true establishment of competence would likely benefit from the addition of an objective assessment (Catton P 1997, Jefferies, Simmons et al. 2007).

The definition of competence varies, but relates to an individual's ability to perform in the workplace to the required standard (Norman 2000). In surgery, competence is required in many areas, including surgical skill and clinical decision-making. At this time, orthopedics is lacking a reproducible, objective measurement of resident competence.

Objective Structured Clinical Examinations (OSCE's) are an important aspect of certifying examinations in many countries. First developed in the late 1970's, an OSCE consists of multiple, timed stations at which each candidate is faced with a discrete patient- or case-based clinical task and is evaluated in an objective and structured way; both the task and the assessment are standardized (Harden and Gleeson 1979, Cohen, Reznick et al. 1990, Newble 2004). An OSCE is especially suited to testing

There is very little published literature on the validity and reliability of OSCE’s in orthopaedics. Recently, Beran et al. used an OSCE to assess history taking and physical examination skills of orthopaedic residents; this testing served to highlight deficiencies in resident knowledge (Beran, Awan et al. 2012). The ability to formulate valid and reliable in-training OSCE’s may become an essential tool in competency-based orthopedic training, and used to assess resident competence regularly throughout training, identify residents with knowledge deficits compared with their peers, and allow appropriate remediation.

A compulsory module at this University is a continuous three-month sports medicine rotation, in which residents are expected to acquire the medical knowledge to deal with common sports medicine conditions. Most of the residents at our institution undertake two sports medicine rotations, one as a junior resident and one as a senior. The majority of exposure to sports medicine occurs during these rotations. We hypothesised that a sports medicine OSCE would demonstrate sufficient reliability and validity to be used for orthopaedic resident in-training assessment.

**METHODS**

**Exam Development**

An established methodology was followed in order to create the OSCE. The 4th edition of Orthopedic Knowledge Update: Sports Medicine was designated as the defined body of knowledge, and assigned as required reading (Kibler 2009). This establishment of a defined curriculum enabled a clear set of learning objectives to be created, with validation by an expert panel (Newble 2004). The majority of the chapters in this text are discussed at resident teaching sessions with a staff surgeon, held on average three times a week - all scenarios tested in the OSCE are covered in this text. A blue print of clinical problems was then created, with a focus on the CanMEDS Role of Medical Expert (Frank 2005). The blueprint utilized six domains of
Medical Expert thought to be pertinent; history, physical examination, image interpretation (investigations and clinical photos), clinical decision-making, consent, and knowledge of surgical techniques.

An hour long OSCE (6 stations) was created - the OSCE was deliberately kept short in order to maximize feasibility. Six stations (each of 10 minutes duration) were written, with each station based on a common clinical scenario. Scenarios included a 51 year old with a traumatic rotator cuff tear, an 18 year old soccer player with a bucket handle meniscus tear, a 21 year old with a CAM lesion of the hip and a labral tear, a 17 year old with a torn anterior cruciate ligament, a 67 year old with a posterior shoulder dislocation, and a 26 year old volleyball player with lateral ligament instability of the ankle.

Computer-based stations were created, with relevant clinical photos, imaging, and intra-operative photographs displayed on screen. At each station, the examiner asked a series of predetermined questions based on the clinical scenario and the images. Candidates’ verbal answers were scored using a binary checklist, assessing performance (mentioned/not mentioned, correct/incorrect). The OSCE questions and binary checklists underwent formal content review at a series of focus groups of orthopaedic surgeons, in order to establish content validity. The knowledge standard was set at the level of the Canadian orthopaedic certifying exam.

Scoring was performed in combination with a single global rating of performance at each station, as it has been shown that global ratings may have a higher degree of reliability and may be more sensitive to levels of training than checklists alone (Reznick, Regehr et al. 1998, Hodges, Regehr et al. 1999, Hodges and McIlroy 2003). In order to increased objectivity, examiners were asked to assign the global rating independently of the checklist score (1 = clear fail, 2 = borderline, 3 = pass, 4 = good pass, 5 = excellent pass) (Hodges, Regehr et al. 1997, Pell G 2006, Pell, Fuller et al. 2010).

The OSCE was trialled on two residents of different postgraduate years. Examiners were also trained in the use and application of the checklist and global ratings. Of the
six examiners, four were part of the focus group, and very familiar with each scenario and the checklist. In regards the examiners who were not part of the focus group, the principal investigator (TD) spent time discussing each scenario, the checklist, and the global rating with these examiners in detail, addressing any queries or concerns. All examiners were instructed that a “pass” on the global rating should be equivalent to a pass at an orthopedic certifying examination level, irrespective of year of training.

**Study Design**

The OSCE was offered to all orthopaedic residents, PGY1 through PGY5, at this institution. Participation was voluntary and written consent was obtained from each resident prior to participation. Examiners at each station were either orthopaedic surgeons or fellows; the same examiner marked each station for the duration of study, conducted over a two-week period. It was not possible to blind all examiners to the year of training of the residents. However, examiners were asked to disregard the year of training when assessing performance.

Information was collected regarding the number of sports medicine rotations each resident had previously undertaken, as well as which residents had undertaken a sports medicine rotation in the last six months. During this period, a competency-based curriculum model had been used, with a set curriculum and exit examinations in the form of OSCE examinations – only a proportion of the residents had been exposed to the new curriculum. After the examination, candidates were invited to provide feedback on the examination via a survey, using a 5-point Likert scale.

**Statistical Analysis**

All data were de-identified, and entered into an Excel spreadsheet, and analysed using SPSS version 19 (Armonk, NY). The checklists, global ratings, postgraduate year of training, sports medicine experience, and questionnaire results were recorded for each resident. Reliability (the precision or reproducibility of the examination process) was calculated using the Alpha Coefficient of Internal Consistency (Cronbach’s Alpha). Individual station reliability scores were calculated using
Cronbach’s Alpha if Item Deleted and Corrected Item-Total Correlation Coefficients (Pell, Fuller et al. 2010). Cronbach’s Alpha if Item Deleted is calculated for the individual stations, by removing each station before recalculating the overall reliability. If removing any one station increases the Cronbach’s Alpha, it implies that the station is performing poorly.

Construct validity was based on the degree to which the OSCE discriminated between different PGY, and assessed using one-way analysis of variance (ANOVA) where year of training was the independent variable and total test scores and domain-specific subscores were the dependant variables. Scheffe’s test was used for post hoc analysis to understand differences in scores between each possible pair of year of training. The correlation between the checklist scores and the global ratings at the individual station level was assessed using Pearson’s correlation and Spearman’s Rho ($R^2$). Concurrent validity was examined by correlating the resident’s performance on the OSCE with their most recent overall and sports component scores on the Orthopedic In-Training Examination (a yearly MCQ examination sat by all orthopaedic residents at this institution).

**Ethical Considerations**

Approval was obtained from the Institutional Research Ethics Boards for this study.

**Funding Source**

No funding was required for the study.

**RESULTS**

A total of 43 residents (64 eligible) participated in the sports medicine OSCE (Table 26). Five of the 43 residents were enrolled in the competency-based curriculum program, while 11/43 residents had undertaken the sports rotation within the last six months.
Table 26. Number of Residents in each Post-Graduate Year (PGY)

<table>
<thead>
<tr>
<th>PGY</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
</tr>
</tbody>
</table>

The total test (OSCE) score was calculated as the average of the 6 station-level (percent correct) scores. The group mean score on the total test was 56.2 (SD 10.4). Mean scores for the 6 stations are presented in Table 27. The highest scoring station was Station 5 (posterior shoulder dislocation), while station 2 (rotator cuff tear) and 3 (CAM lesion of the hip and a labral tear) were the most difficult stations.

Table 27. Group (n=43) mean scores and standard deviation for each of six OSCE stations
The Alpha Coefficient of Internal Consistency was 0.91 for the 6 stations, a result within acceptable limits for a high stakes OSCE (generally accepted as 0.8 or higher) (Cohen, Reznick et al. 1990). Individual station (item) parameters are detailed in Table 28. The Cronbach’s Alpha if Item Deleted decreased for every station except for station 1 (meniscal tear), however this increase was minimal. The Corrected Item Total Correlation Coefficients for the 6 stations were all very high. The $R^2$ was 0.96 overall, ranging between 0.77 and 0.92 for each of the stations, with all stations demonstrating significant correlation between the checklist scores and the global rating ($p<0.01$) on both Pearson Correlation and the non-parametric correlations (Spearman’s Rho).

<table>
<thead>
<tr>
<th>Station</th>
<th>Cronbach’s Alpha if Item Deleted</th>
<th>Corrected Item-Total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Meniscal tear</td>
<td>0.92</td>
<td>0.57</td>
</tr>
<tr>
<td>2. Rotator cuff tear</td>
<td>0.88</td>
<td>0.83</td>
</tr>
<tr>
<td>3. CAM lesion hip</td>
<td>0.90</td>
<td>0.68</td>
</tr>
<tr>
<td>4. ACL rupture</td>
<td>0.88</td>
<td>0.83</td>
</tr>
<tr>
<td>5. Posterior shoulder dislocation</td>
<td>0.89</td>
<td>0.81</td>
</tr>
<tr>
<td>6. Ankle instability</td>
<td>0.88</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Table 28. Individual station parameters

Scores on total test score were significantly higher for PGY5 residents than PGY1, 2 and 3 residents ($p<0.05$); no significant difference was seen between PGY4 and PGY5 (Figure 11). Total checklist scores for PGY4 were significantly increased compared to PGY1 only; no significant difference was seen between PGY1, 2 and 3. A similar pattern of results was seen for the mean of 6 station-level global ratings.
Figure 11. (A) Box plot of total checklist score (y) versus postgraduate year of training (x). Scores on total checklists were significantly higher for PGY5 residents than PGY1, 2 and 3 residents (p<0.05); no significant difference was seen between PGY4 and PGY5 (Figure 2). (B) Box plot of global rating 1 to 5 (x) versus postgraduate year of training (y). A significant difference (p<0.0001) was seen between PGY1 and PGY5 residents. * and * represent group outliers.
Items corresponding to each of six domains were averaged across stations, and Cronbach’s Alpha was calculated for each domain (Table 29). One-way analysis of variance (ANOVA) of the total sub-scores for each of the domains (with the exception of history) showed a statistical difference depending on postgraduate year of training of the resident. Post hoc analyses showed that the domains of clinical decision-making and surgical technique had the greatest discriminatory ability, with the checklist scores of the PGY5 residents significantly different to PGY1, 2 & 3 (Figure 12).

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Number of Items</th>
<th>Cronbach’s Alpha Coefficient</th>
<th>ANOVA for PGY effect</th>
<th>Post Hoc Tests (sig &lt;0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>4</td>
<td>0.49</td>
<td>F₄,₃₈ = 0.61, p = .993</td>
<td>No difference</td>
</tr>
<tr>
<td>Examination</td>
<td>12</td>
<td>0.77</td>
<td>F₄,₃₈ = 6.14, p = .001</td>
<td>PGY4,5 &gt; PGY 1,2,3</td>
</tr>
<tr>
<td>Image Interpretation</td>
<td>24</td>
<td>0.86</td>
<td>F₄,₃₈ = 8.57, p = .000</td>
<td>PGY5 &gt; PGY1,2</td>
</tr>
<tr>
<td>Clinical Decision-Making</td>
<td>20</td>
<td>0.82</td>
<td>F₄,₃₈ = 11.26, p = .000</td>
<td>PGY5 &gt; PGY1,2,3</td>
</tr>
<tr>
<td>Consent</td>
<td>5</td>
<td>0.50</td>
<td>F₄,₃₈ = 3.36, p = .019</td>
<td>PGY5 &gt; PGY1</td>
</tr>
<tr>
<td>Surgical Techniques</td>
<td>6</td>
<td>0.78</td>
<td>F₄,₃₈ = 12.53, p = .000</td>
<td>PGY5 &gt; PGY1,2,3</td>
</tr>
</tbody>
</table>

Table 29. Inter-item reliability of sub-scores and effect of year of training on domains of performance.
Figure 12. Box plots of total checklist scores compared with the postgraduate year of training for each of the six domains. There was no significant difference in the domain of history ($p = 0.993$) (Fig. 2-A), but a significant difference was seen in the domains of examination ($p = 0.001$) (Fig. 2-B), image interpretation ($p < 0.001$) (Fig. 2-C), clinical decision-making ($p < 0.001$) (Fig. 2-D), consent ($p = 0.019$) (Fig. 2-E), and surgical technique ($p < 0.001$) (Fig. 2-F). The degree symbol and the asterisk represent group outliers.
No significant difference was seen in either total checklist scores or global ratings between CBC and regular stream residents. However, a significant effect of the number of sports medicine rotations each resident had undertaken (0, 1, 2 or 3), was seen on total test scores and in all domains with the exception of history (Figure 13). Two-way analysis revealed that this statistically significant increase in total checklist scores was independent of year of training (p<0.05). No significant improvement in total checklist scores or domain scores was seen with residents who had undertaken a sports medicine rotation within the last six months (relative to those with more remote experience), with the exception of history, although a trend was seen for improved total test scores.
Figure 13. Box plots of total checklist scores compared with the number of sports rotations for each of the six domains. Again, there was no significant difference in history (p = 0.281) (Fig. 3-A), but a significant difference was seen in the domains of examination (p = 0.005) (Fig. 3-B), image interpretation (p = 0.001) (Fig. 3-C), clinical decision-making (p = 0.001) (Fig. 3-D), consent (p = 0.01) (Fig. 3-E), and surgical technique (p = 0.001) (Fig. 3-F). The degree symbol represents group outliers.
For the 35 subjects who had written the OITE, total test scores were highly correlated with both OITE total test scores (Pearson correlation coefficient $r = .72, p < 0.001$) and OITE sports subscores ($r = .61, p < 0.001$). High positive correlations were also seen between OSCE domain subscores for physical examination, image interpretation, clinical decision-making, and knowledge of surgical procedures (Table 30).

<table>
<thead>
<tr>
<th>OSCE Domain Subscore</th>
<th>OITE Total Score</th>
<th>OITE Sports Subscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>-.075</td>
<td>.076</td>
</tr>
<tr>
<td>Examination</td>
<td>.585**</td>
<td>.457**</td>
</tr>
<tr>
<td>Image Interpretation</td>
<td>.744**</td>
<td>.627**</td>
</tr>
<tr>
<td>Clinical Decision-Making</td>
<td>.736**</td>
<td>.629**</td>
</tr>
<tr>
<td>Consent</td>
<td>.383*</td>
<td>.282</td>
</tr>
<tr>
<td>Surgical Techniques</td>
<td>.728**</td>
<td>.560**</td>
</tr>
</tbody>
</table>

Table 30. Correlations (Pearson’s r) between OSCE sub-scores and OITE scores for 35 participants. Eight PGY1 residents had not written an OITE at the time of this study, so data was unavailable for correlation. ** Correlation is significant at the 0.01 level (2-tailed) * Correlation is significant at the 0.05 level (2-tailed).

There was a 100% completion rate of the post OSCE survey. Overall, 91% of the residents agreed or strongly agreed that the clinical scenarios accurately reflected scenarios a general orthopedic surgeon might encounter in their practice, while 80% agreed or strongly agreed that an OSCE was the best method of testing their medical knowledge. In regards to usefulness, 100% of residents agreed or strongly agreed that this OSCE helped prepare them for their final exams, with 98% of residents agreeing or strongly agreeing that it would be useful to sit an in-training OSCE after every module or rotation.
DISCUSSION

The result of this study, one of the first to examine the use of an OSCE in orthopedics, has revealed excellent reliability. Furthermore, this OSCE has shown good validity, with the ability to discriminate between different years of training, both overall and within domains of knowledge. Importantly, it has been demonstrated that it is possible to achieve these goals with a one-hour OSCE, and without the use of standardized patients, making this a feasible assessment tool. We believe that by following the standard process outlined in Figure 1, reliable and valid OSCEs can be developed for orthopaedic trainee evaluation, in all modules of orthopaedics.

The overall Cronbach’s Alpha of 0.91 in this OSCE was extremely high; as a rule greater than 0.6 is required for making acceptance decisions (i.e. selection to resident training), with 0.80 or greater required for making pass-fail decisions in high stakes examinations (Cohen, Reznick et al. 1990). It is generally accepted that examinations with low numbers of stations are more likely to have lower reliabilities; an OSCE used in general surgery, examining multiple subspecialties, utilized 38 stations (Cohen, Reznick et al. 1990). This latter type of assessment model would not be feasible for regular in-training assessment of competence. It may be that testing only a narrow range of content (sports medicine) is responsible for the high reliability seen.

Some information has been obtained regarding knowledge domains that improve over postgraduate training. In all domains except history taking, the PGY5 residents demonstrated statistically superior checklist scores. In this study, the PGY5 residents sat the sports medicine OSCE just before sitting their final certifying examinations – their superiority in the domains of surgical technique and clinical decision-making is in part expected. History was a poor discriminator – it may be that this is taught well early in the medical education process. However, the scores are not particularly high across all years – in fact history taking may need increased attention in the curriculum. The consenting process seemed poorly understood across all post-graduate years. This provided us with valuable feedback; in future rotations an increased emphasis will be placed on consent and postoperative complications.
We found that increasing sports medicine experience correlated to an increase in scores. Naturally, increasing experience is closely related to advanced year of training, however, further analysis suggested that this effect was independent of year of training. Surprisingly, having undertaken a recent sports medicine rotation (run along a competency-based model) did not have a significant effect on performance, although there was a trend to improved scores. This assessment, however, was limited in our study due to the low number of residents (11) who had been exposed to the new curriculum. We also identified a correlation between the results of the sports medicine OSCE, with the OITE total test scores and the OITE sports subscores. This correlation may be explained by improved sports medicine knowledge enhancing performance in both the OSCE and the OITE, or by the inherent test-taking abilities of some residents. Obviously, good results on each form of assessment signify sound medical knowledge and judgement, and do not imply competence in surgical skill.

At this time, conclusions regarding competence cannot be drawn, as some of the residents yet to undertake a sports medicine rotation. These residents scored poorly on the OSCE, as expected. In order to address the issue of standard setting, evaluation of residents can only take place upon completion of a module. This research is currently underway at this institution, focussing on an absolute or criterion-referenced approach, in which the pass/fail point is set on the basis of what a student should know (Turnbull 1989, Norcini 1999, Newble 2004, Schoonheim-Klein, Muijtjens et al. 2009). More than 30 different criterion-referenced methods have been described – our research will focus on finding the most reliable and feasible method (Cusimano 1996). Furthermore, the use of global ratings for each of the domains, and the use of fellows rather than staff surgeons will be studied.

It is uncertain whether a PGY1 resident should have the same expected degree of competence as a PGY5 resident after a rotation. We feel that while it might not be reasonable to expect the same level of surgical skill from a junior resident, achieving competence in clinical judgement and decision-making should be possible. This is of course, very debatable – it could be argued that increased experience in orthopaedics will improve performance in both medical knowledge and surgical skill. The results of
this study seem to suggest that increasing level of experience is correlated to improved performance (especially image interpretation, clinical decision-making and surgical technique) more so than exposure to a recent sports medicine rotation, although this was not the primary goal of the research. Again, we hope to investigate this issue further by studying methods of standard setting, especially in relation to determining the competence of junior versus senior residents.

This sports medicine OSCE was designed to test medical knowledge and clinical judgment, not to assess the acquisition of surgical skills such as arthroscopy. Options for assessing surgical skills include Objective Structured Assessment of Technical Skills (OSATS), Structured Assessment of Clinical Evaluation Report (STACER), and components of the ITER. At our institution, after bimonthly teaching in a dry arthroscopy laboratory, residents are assessed with an OSATS (global ratings) while performing an ACL reconstruction, and tying an arthroscopic knot on sawbone models. Research is being conducted to determine if there is a correlation between the OSATS results and the sports medicine OSCE, with a goal to develop a comprehensive assessment model for residents at the end of rotations.

We recognize that the definition of an OSCE is variable (Newble 2004). An OSCE is an organizational framework, consisting of multiple, timed stations at which each candidate is faced with a discrete clinical task and is evaluated in an objective, structured way (Harden and Gleeson 1979, Cohen, Reznick et al. 1990, Newble 2004). Stations may or may not involve a standardized patient (SP) (Hodges, Regehr et al. 1999). One of the major advantages of not using standardized patients is that it significantly reduces running costs to a minimum, an important factor if an OSCE is to be used on a regular basis. If desired, the structure of the OSCE could be altered such that examiners acted as a SP, answering history questions posed by candidates.

In this study, we used a combination of checklists and global ratings. Regehr et al. demonstrated that global ratings completed without the use of a task-specific checklist, were equally valid and reliable as global ratings used in association with a checklist, suggesting that completion of a checklist does not influence the use of a global rating scale (Regehr, MacRae et al. 1998). Furthermore, the global rating scales
demonstrated greater interstation reliability, construct validity, and better concurrent validity than did the checklists “(CME)”. It is thought that relying on checklists alone results in possible over-objectification of performance, and that global ratings can be used by expert raters to assess the less quantifiable aspects of performance (Norman, Van der Vleuten et al. 1991, Van der Vleuten, Norman et al. 1991). For this reason, a combination was used.

A potential issue in our methodology was that the PGY level of some residents will have been known to examiners, raising the potential for bias. Having residents rotate from station to station, and examiner to examiner reduces this risk. Examiners were also asked to disregard the PGY level of the resident when assigning the global rating. While using examiners from outside institutions could reduce this potential for bias, this would inevitably raise costs and reduce feasibility. It is planned to extend the OSCE assessment to the arthroplasty module at our institution, and increase the use of orthopaedic fellows rather than staff as examiners, however the potential for bias remains. Ultimately, by conducting formal OSCE assessment at the conclusion of every module in training, we believe that this potential for bias will be minimized. This would be the ideal model for implementation on a national basis.

Other limitations of this study include the relatively low number of candidates, especially compared to the published literature on the use of OSCEs in medical school examinations. However, as this is the largest training program in the Canada, it would be difficult to increase numbers without using multiple centres. It may also be there was some selection bias involved, as residents were asked to volunteer to sit the OSCE; some poorly performing residents may not have participated, potentially skewing our results. Furthermore, this OSCE was run over a period of two weeks; ideally the OSCE would be administered in a single setting, in order to prevent resident awareness of exam cases. However evidence in the literature seems to show that information sharing amongst students has little effect on overall mean scores, especially in the absence of access to the checklists (Colliver, Travis et al. 1992, Swartz, Colliver et al. 1993, Champlain, Macmillan et al. 2000). Finally, this OSCE does not assess surgical skill, only medical knowledge and judgment.
CONCLUSION

It has been demonstrated that a sports medicine OSCE can have acceptable reliability and validity, sufficient to be used to assess the in-training competency of orthopedic residents. Furthermore, this examination method is feasible, with minimal costs and high acceptability among candidates.
APPENDIX D. Simulation of anterior cruciate ligament reconstruction (ACLR) in a dry model
T Dwyer, J Slade Shantz, J Chahal, D Wasserstein, R Schaschar, B Devitt, J Theodoropoulos, C Ringsted, D Ogilvie-Harris

ABSTRACT

Background
As the demand increases for demonstration of competence in surgical skill, the need for validated assessment tools also increases.

Hypothesis / Purpose
The purpose of this study was to validate a Sawbones model for the assessment of performance of anterior cruciate reconstruction (ACLR). We hypothesized that the combination of a checklist and a previously validated global rating scale would be a valid and reliable means of assessing ACLR when performed by residents in a dry model.

Method
All residents, sports medicine staff and fellows were invited to perform a hamstring ACLR, using anteromedial drilling, and endobutton fixation on an ACL Sawbones model. Previous exposure to knee arthroscopy and ACLR was recorded. A detailed surgical manuscript and technique video was sent to all participants prior to the study. Residents were evaluated by staff surgeons using task-specific checklists created using a modified Delphi procedure, and the Arthroscopic Surgical Skill Evaluation Tool (ASSET). Each procedure (hand movements and arthroscopic video) was recorded and scored by a fellow blinded to the year of training of each participant.
Results

A total of 29 residents, six fellows and five staff surgeons performed an ACLR on the sawbones model (40 total). The internal reliability (Cronbach’s Alpha) of the test using the total ASSET score was very high (>0.9). One-way analysis of variance for the total ASSET score and the total checklist score demonstrated a difference between participants based upon year of training (p<0.001). Post hoc analysis demonstrated a significant difference in global ratings and checklist scores between junior residents (PGY1-3) and senior residents (PGY4&5), senior residents and fellows, and fellows and staff (p<0.05). A good correlation was seen between the total ASSET score and prior exposure to knee arthroscopy (0.73) and ACLR (0.65). The inter-rater reliability (ICC) between the examiner ratings and the blinded assessor ratings for the total ASSET score was very high (>0.8).

Conclusion

The results of this study provide evidence that the performance of an ACLR in a Sawbones model is a reliable method of assessing resident’s knowledge of the steps and instrumentation required, and shows evidence of validity.

INTRODUCTION

As the demand increases for demonstration of competence in surgical skill in postgraduate training, the need for validated assessment tools has also increased. Lately, attention has been focused on validating assessment forms for knee arthroscopy (Elliott, Caprise et al. 2012, Koehler, Amsdell et al. 2013, Slade Shantz, Leiter et al. 2013), partial meniscectomy (Insel, Carofino et al. 2009), meniscal repair (Alvand, Logishetty et al. 2013), and shoulder arthroscopy (Hoyle 2012).

In 2001, Farnworth et al. identified that an anterior cruciate ligament reconstruction (ACLR) performed by a resident in the operating room took an average of 42 minutes longer, and cost an additional $900 than if the procedure was performed by a staff surgeon (Farnworth, Lemay et al. 2001). While intra-operative performance of surgical procedures by a trainee cannot be replaced, surgical simulation offers the opportunity
for residents to master the ability to perform the steps of complex operations, and develop familiarity with instrumentation prior to the operating room.

Options for skills training and assessment in arthroscopy include the use of cadavers (Insel, Carofino et al. 2009, Tashiro, Miura et al. 2009, Elliott, Caprise et al. 2012, Koehler, Amsdell et al. 2013, Slade Shantz, Leiter et al. 2013), dry models (Howells, Gill et al. 2008, Pollard, Khan et al. 2012, Alvand, Logishetty et al. 2013) and virtual reality (Gomoll, Pappas et al. 2008, Cannon 2014) - each option has its own particular advantages and disadvantages. While it could be argued that high-fidelity models such as cadavers are the gold standard, the associated costs are extremely high, and thus not well suited to regular teaching and assessment of technical skill. Virtual reality simulators hold promise, but at this time the ability to perform complex procedures such as ACLR is limited (Cannon 2014).

Dry models, such as Sawbones, offer advantages including reduced cost, ready availability, and simple laboratory setup (in comparison to cadavers in wet laboratories). Previous work has demonstrated that training on dry models improves performance of diagnostic knee arthroscopy on both cadavers (Butler, Olson et al. 2013) and in the operating room (Howells, Gill et al. 2008). It has also been shown that, in a model of ulna fracture fixation, a synthetic model was superior to a virtual reality simulator (LeBlanc, Hutchison et al. 2013).

In 2013, Koehler et al. published the Arthroscopic Surgical Skill Evaluation Tool (ASSET), a global rating scale with eight domains, including camera and instrument dexterity, flow of the procedure and quality of the procedure (Koehler, Amsdell et al. 2013). Each domain is rated from 1 to 5, based upon the Drefus model of skill acquisition (Novice, Competent, Expert). While the ASSET at this time has only been validated in the setting of knee arthroscopy on cadavers, the ASSET was designed to be generalizable to multiple arthroscopic procedures.

The purpose of this study was to validate the use of a Sawbones model for the assessment of performance of ACLR by residents. We hypothesized that the
combination of task-specific checklists and the ASSET would be a valid and reliable means of assessing ACLR when performed in a low-fidelity Sawbones model.

METHODS

All residents in our training program were offered the opportunity to perform an ACLR on an ACL Sawbones model (Sawbones, Washington, United States), utilizing an encapsulated knee insert (model 1414-1) inside a soft tissue with skin (model 1413-1) (Figure 1). Standard 30° arthroscopic cameras with high definition video systems were used for the procedure, in association with ACLR equipment including drill guides, beath pins, tibial and femoral reamers, as well as interference and suspensory fixation (Smith & Nephew, Andover, MA).
Prior to performing the procedure, each participant was emailed a list of steps required to perform a hamstring ACLR, using an anteromedial femoral drilling technique, with endobutton fixation (Smith & Nephew, Andover, MA) on the femoral side, and interference screw on the tibial side. The list of steps for completion of the technical aspects of the procedure was also available on the day during the procedure. An instructional video was also provided for review prior to assessment. All accessory anteromedial portals were premade in the soft tissue skins of the sawbone models, all hamstring grafts were 8 mm in diameter, and all endobuttons had a 20 mm loop.

Figure 14. A: Equipment provided for ACL reconstruction in Sawbones model. B: View of the Sawbones model, with the accessory anteromedial portal being used in conjunction with an offset guide to drill the femoral tunnel. C: Drilling of the tibial tunnel. D: Passage of the femoral beath pin using an offset guide.
Residents were evaluated using task-specific checklists (Table 31 – 33), and the ASSET. The ACLR task-specific checklists were created using review by experienced staff surgeons – consensus was achieved using a modified Delphi procedure conducted by way of multiple electronic surveys, using a previously described methodology (Koehler, Amsdell et al. 2013). The examiners (staff and fellows) were instructed to disregard year of training, and to rate each participant as competent if able to perform at the level of a practicing orthopedic surgeon.

<table>
<thead>
<tr>
<th>Task</th>
<th>Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uses anteromedial portal</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Correct knee position for technique i.e. hyperflexed</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Uses femoral offset guide</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Appropriate femoral offset guide i.e. 6 mm</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Correct position on wall i.e. between 10 and 11</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Estimates guide pin exit before drilling</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Correct guide pin exit i.e. between PFJ and lateral intramuscular septum</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Passes endobutton drill</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Measures depth of Endobutton tunnel</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Drills with 8 mm femoral reamer to correct depth (i.e. within 10 mm)</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Checks back wall</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Back wall intact</td>
<td>□ Yes □ No</td>
</tr>
</tbody>
</table>

Table 31. Task-specific checklist for the drilling of the femoral tunnel, using an anteromedial technique, for an 8 mm hamstring graft with fixation using a 20 mm loop Endobutton.
<table>
<thead>
<tr>
<th>Task</th>
<th>Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knee at 90 degrees of flexion</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Sets tibial jig angle to 55 degrees</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Correct application of anteroposterior landmarks for tibial entry point i.e. posterior border of lateral meniscus</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Places tibial jig centrally with regards to medial – lateral position</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Guide pin entry in center of tibial metaphysis</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Drills with 8 mm tibial reamer</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>Care of guide pin when drilling</td>
<td>□ Yes □ No</td>
</tr>
</tbody>
</table>

Table 32. Task-specific checklist for the drilling of the tibial tunnel.
In order to determine which participants were competent, each participant was also given a final five-point overall global rating scale (GRS) corresponding to the Drefus model of skill acquisition (novice, advanced beginner, competent, proficient, expert) (Batalden, Leach et al. 2002, Carraccio, Benson et al. 2008). Each participant was determined to be competent if they achieved an overall GRS of competent or above, or if they scored a minimum of 3 in each of the 8 ASSET domains as per the study by Koehler et al (Koehler, Amsdell et al. 2013). In a third method, the number of participants reaching a minimal level of competence was also assessed using an ASSET score of 24 or greater.

Each procedure was recorded, with videotaping of the hand movements and instrument usage, and arthroscopic video recordings of the intra-articular procedure. These videos were scored by a single orthopedic sports medicine fellow, blinded to
the year of training of each resident, using the checklists and the ASSET. The blinded fellow used the checklist as a guide - despite the use of videotaped hand movements, many components of the checklist (such as use of the anteromedial portal, estimating guide pin exit, setting of the tibial jig angle) could not be consistently scored by the blinded reviewer, and were thus not used in the analysis.

Examiners provided assistance throughout the case as directed by the resident, but no verbal instruction was allowed. The number of sports rotations (three months each) each resident had previously undertaken was recorded - residents were also asked to estimate their prior exposure to knee arthroscopy (number of cases) and ACLR procedures (number). Costs for equipment and models were also calculated per participant.

**Statistical analysis**

STATA version 13 (College Station, Texas) was used to perform all analyses. The total score of the ASSET was calculated as the sum of the eight domains, with a maximum score of 38 (Koehler, Amsdell et al. 2013). The overall reliability was assessed using Cronbach’s Alpha. Correlation between the total checklist score and the ASSET was sought using Pearson’s correlation. Construct validity (the ability to discriminate between different postgraduate year of training) was assessed using one-way analysis of variance (ANOVA), with year of training as the independent variable. Concurrent validity was sought seeking association between total ASSET score, the number of previous sports rotation, and the previous knee arthroscopy and ACLR experience. Inter-rater reliability (using faculty rating and the blinded assessor ratings) was calculated using the intraclass correlation coefficient (ICC) for both the total ASSET score and for the total checklist score.

**Ethics**

Approval for this study was provided by the institutional research ethics board prior to commencement.
RESULTS

Over two days, 29 residents performed an ACLR on the Sawbones model. Five orthopedic sports medicine fellows and six staff surgeons also performed the ACLR, for a total of 40 participants. The internal consistency / reliability (Cronbach’s alpha) of the test using the total ASSET score was very high (>0.9). The reliability for the total femoral checklist score was 0.75, for the tibial checklist score was 0.78, and 0.68 for the graft passage and fixation checklist. A good correlation was seen between the total ASSET score and the total checklist score (0.71).

The inter-rater reliability between the examiners ratings, and the blinded assessor rating for the total ASSET score using ICC was 0.85 (95% CI 0.71 – 0.98). With regards to mean percent correct of the total checklist score, this was 68.5% for junior residents, 77.6% for senior residents, 87.3% for fellows, and 93.2% for staff surgeons.

One-way analysis of variance demonstrated a significant difference between participants based upon year of training for the total ASSET score (p<0.001), and for the total checklist scores (p<0.001) (Figure 15). Unpaired t-test demonstrated a significant difference between junior (postgraduate year (PGY) 1,2,3) and senior (PGY 4&5) residents (p<0.001) for the total ASSET score – there was also a significant difference between junior and senior residents for the total checklist score (p=0.025). There was also a significant difference between senior residents and fellows on the ASSET global rating (p=0.005), and on the checklist scores (p=0.006). A significant difference was seen between the fellows and staff surgeons on the ASSET global rating (p=0.03), but not on the checklist scores (p=0.25). The number of participants in each group deemed minimally competent by each of the three methods is see in Table 34.
Figure 15. A: Boxplot for the total ASSET score for junior residents (postgraduate year 1-3), senior residents (postgraduate year 4&5), fellows, and staff surgeons. A significant difference was seen between junior and senior residents, senior residents and fellows, and fellows and staff surgeons. B: Boxplots for the total checklist score for the four groups of participants. A significant difference was also seen between junior and senior residents, as well as senior residents and fellows – no significant difference was seen between fellows and staff surgeons.
<table>
<thead>
<tr>
<th>Year of Training</th>
<th>Number</th>
<th>Number competent or above by overall GRS</th>
<th>Number competent by ASSET score (minimum of 3 in each domain)</th>
<th>Number competent by ASSET score (≥ 24/38)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGY1</td>
<td>5</td>
<td>0/5 (0%)</td>
<td>0/5 (0%)</td>
<td>0/5 (0%)</td>
</tr>
<tr>
<td>PGY2</td>
<td>8</td>
<td>1/8 (12.5%)</td>
<td>1/8 (12.5%)</td>
<td>1/8 (12.5%)</td>
</tr>
<tr>
<td>PGY3</td>
<td>6</td>
<td>1/6 (16.7%)</td>
<td>1/6 (16/7%)</td>
<td>1/6 (16.7%)</td>
</tr>
<tr>
<td>PGY4</td>
<td>5</td>
<td>4/5 (80%)</td>
<td>3/5 (60%)</td>
<td>5/5 (100%)</td>
</tr>
<tr>
<td>PGY5</td>
<td>5</td>
<td>5/5 (100%)</td>
<td>5/5 (100%)</td>
<td>5/5 (100%)</td>
</tr>
<tr>
<td>Fellows</td>
<td>5</td>
<td>5/5 (100%)</td>
<td>5/5 (100%)</td>
<td>5/5 (100%)</td>
</tr>
<tr>
<td>Faculty</td>
<td>6</td>
<td>6/6 (100%)</td>
<td>6/6 (100%)</td>
<td>6/6 (100%)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>40</td>
<td>22/40 (55%)</td>
<td>21/40 (53%)</td>
<td>22/40 (55%)</td>
</tr>
</tbody>
</table>

Table 34. Number of participants in the study and their year of training. Three different methods of determining a minimal level of competency are shown. PGY – postgraduate year of training. GRS – global rating scale.
With regards to video review analysis, one-way analysis of variance demonstrated a significant difference between participants based upon year of training for the total ASSET score \((p<0.001)\) (Figure 16). Unpaired t-test demonstrated a significant difference between junior and senior residents \((p=0.01)\) for the total ASSET score, between senior residents and fellows \((p<0.001)\), and between the fellows and staff surgeons on the ASSET global rating \((p=0.004)\).

Figure 16. Boxplot for the total ASSET score for all participants, based upon blinded video review analysis. Again, a significant different was seen between participants based upon year of training for the total ASSET score \((p<0.001)\), as well as between all groups

The correlation between the total ASSET score for the number of previous sports rotations was 0.58, 0.73 for the number of knee arthroscopies, and 0.65 for the number of ACLR (Table 35). The ACLR study was conducted using 40 ACL Sawbones ($69.75 each), four replacement skins ($83.70 each), and four hamstrings grafts ($11.30). All instrumentation including drills, screws, and endobuttons were donated, and reused throughout the study. The total cost was $3,170.00, which amounted to $79.25 per participant.
Table 35. Details of the sports medicine experience and exposure of each group of participants in this study.

**DISCUSSION**

The results of this study demonstrate that performance of an ACLR in a Sawbones model is a reliable method of assessing resident’s knowledge of the steps and instrumentation required, and shows some evidence of validity. These findings allow us to ensure residents and fellows are competent to perform this procedure in the simulation setting, prior to performance in the operating room, with implications for both patient safety and cost.

A survey of orthopedic program directors in the USA identified that 76% of programs had a surgical skills laboratory, with 80% agreeing that surgical skills simulations should become a required part of training (Karam, Pedowitz et al. 2013). A significant issue raised by program directors was cost - a systematic review of simulation studies in 2013 identified that the majority of studies do not estimate cost (Zendejas, Wang et al. 2013). In this study we have shown that it costs approximately $80 to simulate an ACLR in Sawbones, much cheaper than in cadaveric models. Furthermore, these
models can be reused, allowing residents and fellows to practice meniscectomy, meniscal repair, and even PCL reconstruction.

A meta-analysis conducted by Cook et al. in 2011 concluded that simulation based medical education improved learner outcomes, compared with no intervention (Cook, Hatala et al. 2011). Overall, the majority of comparative studies in the surgical specialties have demonstrated that low fidelity simulators are similarly effective, but less expensive than high fidelity simulators with regards to the acquisition of surgical skill (Anastakis, Regehr et al. 1999, Matsumoto, Hamstra et al. 2002, Grober, Hamstra et al. 2004, McDougall, Kolla et al. 2009, LeBlanc, Hutchison et al. 2013). While virtual reality simulators are extremely encouraging, at this time they are used primarily to allow the development of arthroscopic localization and probing skills (Gomoll, Pappas et al. 2008), with limited ability at this time to simulate more complex procedures common in sports medicine (Cannon, Nicandri et al. 2014). For this reason we believe that the use of Sawbones models is a valuable component of simulation in sports medicine, because they allow performance of complex procedures using identical instrumentation as live surgery.

Previous randomized trials have demonstrated that training on virtual reality simulators and bench-top simulators improves the performance of knee arthroscopy on live patients (Cannon 2014), diagnostic knee arthroscopy on cadavers (Butler, Olson et al. 2013), and in the operating room (Howells, Gill et al. 2008), and of shoulder arthroscopy in cadavers (Henn, Shah et al. 2013). However, it remains to be seen if the results of this study will correlate with the performance of ACLR in the operating room.

While we have demonstrated that the Sawbones model can distinguish between participants of varying experience (year of training, knee arthroscopic and ACLR experience, the number of previous sports rotations), an important component of validity evidence (the degree to which a study accurately reflects or assesses the specific concept that the researcher is attempting to measure) would be correlation with performance of an ACLR in cadavers and/or patients. While this is an important next step of research, the assessment of performance in cadavers is expensive, and
the assessment of performance in the operating room has numerous challenges, including difficulty standardizing operations (Reznick, Regehr et al. 1997), issues of patient safety, as well as time restraints (Farnworth, Lemay et al. 2001). As such, we believe that simulation-based assessment offers part of the solution to this complex problem, allowing a minimal level of competence to be demonstrated by trainees prior to performance in the operating room.

There are obviously many ways to perform an ACLR, and by no means do we claim that the hamstring ACLR using anteromedial drilling of the femoral tunnel is the gold standard, nor that instruments such as femoral offset guides are mandatory. This particular method of ACLR was chosen because it is a common technique that residents are exposed to at our institution. We also specified in the list of steps provided for each resident what angle to set the tibial jig at, and what length of loop the endobutton would have, in order to limit variability in the procedure between participants. However, we believe the generalizability of the ASSET means it could easily be used to assess performance of bone patella tendon bone ACLR using transtibial drilling for example, in conjunction with newly created task specific checklists, or using modifications of our checklists.

Checklists also have some limitations. For example, in the tibial tunnel location we ask the examiner to ensure the participant is considering both antero-posterior and medial-lateral landmarks, but we do not list specifics for two reasons. Firstly, one of the limitations of the ACL Sawbones model is there are few landmarks in the notch, especially on the tibia. Secondly, we use this model is to allow trainees to become familiar with the steps of the operation and the instruments – the finesse of tibial tunnel landmarks is best gained in the operating room.

Review of the checklist scores for each of the groups of participants showed a tendency for some common deficiencies. For example, many of the residents and some of the fellows failed to estimate femoral guide pin exit prior to drilling, a useful technique that ensures maximal hyperflexion has been achieved. Furthermore, the majority of junior and senior residents did not actively ensure the integrity of the back wall after drilling of the femoral tunnel – while this is not critical in suspensory fixation
reconstruction, it certainly is when using screw fixation. Junior residents were also less likely to ensure the endobutton had flipped after passage, and would often not tension the hamstring graft before securing it with the tibial screw.

In this study the ASSET was used as the global rating scale, rather than some of the other tools recently published in the literature (Insel, Carofino et al. 2009, Slade Shantz, Leiter et al. 2013). While we use the OSATS global rating scale to assess simulated open procedures at our institution (Martin, Regehr et al. 1997), we chose to use the ASSET both because it was designed to be generalizable for all arthroscopic procedures, and because it is anchored using the Drefus model of skill acquisition (Koehler, Amsdell et al. 2013). We find that these descriptors are very easy for faculty and residents alike to understand, and relevant in the setting of Competency Based Medical Education. While the ASSET was originally validated using only the intra-articular portion of knee arthroscopy, in this study we included review of hand motions – the domains of instrument dexterity and bimanual dexterity seemed particularly useful in this setting.

Moving forward, simulation studies will most likely be used to ensure residents and fellows demonstrate a minimal level of competence in the performance of technical procedures prior to performance in the operating room. How this standard is set is therefore important. In their original paper studying the use of the ASSET score in knee arthroscopy, Koehler et al. believed that a minimum score of three in each of the eight domains was required to assign a grade of competent (Koehler, Amsdell et al. 2013). It is important to note that, in our study, faculty were not aware that giving a score of two in any domain would designate a resident as not competent. Interestingly, minimal difference in pass rates was seen whether we used this method, a minimum ASSET score of 24, or an overall GRS of competent or above. Despite a wide standard deviation, all of the fellows were able to achieve a minimal level of competency, with all of them being given an overall GRS of proficient or more. We also encountered a wide standard deviation for the senior residents – all PGY5 residents were deemed competent by all methods, as opposed to the PGY4 residents, some of whom failed to achieve a minimal level of competence.
The limitations of this study are that only one blinded reviewer was used, as well as only one rater for each participant. However the study by Koehler demonstrated excellent inter-rater reliability, and our study used a blinded reviewer to eliminate the possibility of bias (e.g., rating participants based upon their level of training). We nominally divided the residents into junior (PGY1-3) and senior (PGY4&5) – we did so because the residents at our institution undertake the sports rotation twice as juniors and seniors. Furthermore, since we used Sawbones models, we were unable to assess important aspects of the procedure such as the harvesting of grafts for ACLR. Another limitation is that it is unclear how many of the residents watched the video or studied the instruction sheets prior to performing the ACLR, which may affect their performance, although possibly as much as their level of experience or previous exposure to ACLR. While a good correlation was seen between previous arthroscopy exposure and performance, log books were not used, and there may be an element of recall bias. Finally, as previously mentioned, the translation of skills attained performing ALCR on a Sawbones into the operating room is unknown at this time, and will require further investigation.

CONCLUSION

The results of this study provide evidence that the performance of an ACLR in a Sawbones model is a reliable method of assessing resident’s knowledge of the steps and instrumentation required, and shows evidence of validity. These models may be used to ensure a minimal level of competence prior to resident and fellow performance of ACLR in the operating room.
REFERENCES

(2007) "ACGME Outcome Project. Common program requirements: general competencies.".
ACGME "Surgery program requirements."
ACGME (2013) "The Orthopaedic Surgery Milestone Project."


the financial impact of orthopaedic surgical training in the operating room." Iowa Orthop J 21: 31-35.


RCPSC (2014) "Competency by Design: Reshaping Canadian Medical Education."


Copyright Acknowledgements

Competency-Based Medical Education: Can Both Junior and Senior Residents Achieve Competence After a Sports Medicine Training Module?
Published by the Journal of Bone and Joint Surgery America, 2015 Dec 2 97(23):1985-91
T Dwyer, S Wright, KM Kulasegaram, J Theodoropoulos, J Chahal, D Wasserstein, C Ringsted, B Hodges, D Ogilvie-Harris

How to set the bar in Competency-Based Medical Education: Standard setting using an Objective Structured Clinical Examination (OSCE)
Published by BMC Medical Education, 2016 Jan 4; 16(1): 1.
T Dwyer, S Wright, KM Kulasegaram, J Theodoropoulos, J Chahal, D Wasserstein, C Ringsted, B Hodges, D Ogilvie-Harris

How to assess Communication, Professionalism, Collaboration and the other Intrinsic CanMEDS Roles in Orthopaedic Residents: Use of an Objective Structured Clinical Assessment (OSCE).
T Dwyer, S Glover Takahashi, M Kennedy Hynes, J Herold, D Wasserstein, M Nousiainen, P Ferguson, V Wadey, M Murnaghan, T Leroux, J Semple, B Hodges, D Ogilvie-Harris

Cognitive and Psychomotor Entrustable Professional Activities: Can Simulators Help Assess Competency in Trainees?
Published by Clinical Orthopedics and Related Research, September 22, 2015
T Dwyer, V Wadey, D Archibald, W Kraemer, J Slade Shantz, J Townley, D Ogilvie-Harris, M Petrera, P Ferguson, M Nousiainen.

Assessing Competence of Orthopaedic Residents: The Reliability and Validity of an Objective Structured Clinical Examination After a Sports Medicine Rotation
Published by the Journal of Bone and Joint Surgery America, 2013 Nov 20;95(22):e177
T Dwyer, J Theodoropoulos, J Herold, P Henry, D Wasserstein, M Murnaghan, V Wadey, B Hodges, J Semple, D Ogilvie-Harris

Simulation of Anterior Cruciate Ligament Reconstruction (ACLR) in a Dry Model
Published by the American Journal of Sports Medicine, 2015 Dec; 43(12): 2997-3004.
T Dwyer, J Slade Shantz, J Chahal, D Wasserstein, R Schachar, KM Kulasegaram, J Theodoropoulos, R Greben, D Ogilvie-Harris