

The Development of a Workplace-Based Surgical Clinic Assessment Tool

Janelle Rekman

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the Master's of Arts in Education degree

Department of Education
Faculty of Education
University of Ottawa

Table of Contents

ABSTRACT	V
LIST OF FIGURES	VI
LIST OF TABLES	VI
LIST OF ACRONYMS & ABBREVIATIONS	VII
SECTION 1: INTRODUCTION TO THESIS	1
PROBLEM STATEMENT	1
STUDY OBJECTIVES/PROJECT PURPOSE	1
<i>Thesis question.</i>	1
<i>Sub-objectives.</i>	1
FUTURE DIRECTIONS	2
SECTION 2: BACKGROUND	3
INTRODUCTION	3
THEORETICAL FRAMEWORK	4
MEDICAL EDUCATION CONTEXT: COMPETENCY BASED MEDICAL EDUCATION.	5
<i>CanMEDS Competencies.</i>	6
<i>Entrustable professional activities and milestones.</i>	7
<i>Entrustability Decision Making and Construct Aligned Scales.</i>	7
Construct aligned scales.	9
Entrustability scales.	9
SURGICAL EDUCATION	10
<i>Brief introduction to surgical education.</i>	10
<i>Surgical assessment context.</i>	10
<i>Workplace-based assessment in surgery.</i>	12
Biases in surgical workplace-based assessment.	13
<i>Workplace-based assessment options in the literature.</i>	15
Summary of workplace based assessment options in the literature.	18
VALIDITY EVIDENCE.....	18
<i>Collecting validity theory evidence.</i>	18
Content evidence.	19
Internal structure evidence.	19
Relations with other variables evidence.	19
Response process evidence.	20
Consequential evidence.....	20
SUMMARY OF BACKGROUND.....	20
SECTION 3: METHODOLOGY	22
SUMMARY - INTRODUCTION	22
<i>Research ethics board approval.</i>	23
<i>Grant support.</i>	23
STEPS OF TOOL DEVELOPMENT	23
<i>Step 1: Determining the purpose of the assessment tool.</i>	23
<i>Step 2: Identifying construct of interest & content validity.</i>	24
Literature review.	24
<i>Step 3: Review with content experts.</i>	24
Introduction.	24
Steps used in the consensus group process.	25
Consensus group one.....	26
Consensus group two.	28
National external expert survey.....	28
<i>Step 4: Item & anchor development.</i>	29
Pre-pilot project.....	29
<i>Step 5: Pilot project</i>	30

Identify participants.....	30
<i>Step 6: Rater orientation sessions.</i>	30
Consent process.....	31
Pre-survey question.....	31
Six-month pilot study.....	31
Psychometric data analysis.....	32
Analysis of short answer questions.....	34
<i>Step 6B: Feasibility assessment - Feedback sessions & survey.</i>	34
Feedback survey.....	35
SECTION 4: RESULTS	36
INTRODUCTION	36
STEPS OF TOOL DEVELOPMENT	37
<i>Step 1: Determine the purpose of the tool.</i>	37
<i>Step 2: Content validity evidence.</i>	37
Literature review (construct of interest).....	37
Identify stakeholders.....	37
<i>Step 3: Review with content experts.</i>	38
<i>Content validity evidence collected.</i>	38
Results of consensus groups.....	38
Summary of consensus group one.....	38
Summary of consensus group two.....	40
External surgical education expert survey.....	42
<i>Step 4: Item development.</i>	47
Global assessment item.....	48
Pre-pilot project.....	48
<i>Step 5: Pilot test the instrument.</i>	48
Pre-survey of residents.....	48
Pilot study descriptive statistics.....	50
Psychometric analysis.....	51
Item analysis.....	51
Generalizability analysis.....	53
Subscale question 12 (Q12) & 13 (Q13) (technical skills).....	55
Global assessment item.....	56
Other analyses.....	57
Short answer question analysis.....	60
Question 14 – Specific aspect of clinic done well.....	60
Question 15 – Specific suggestion for improvement.....	60
<i>Step 6B: Feasibility check—length, clarity, and usability.</i>	61
Informal feedback throughout the pilot project.....	61
Feedback sessions.....	62
Staff feedback session.....	62
Resident feedback session.....	67
SECTION 5: DISCUSSION	70
INTRODUCTION	70
CONTENT EVIDENCE	70
<i>Pertaining to purpose.</i>	70
<i>Pertaining to content.</i>	71
INTERNAL STRUCTURE EVIDENCE	75
<i>Item analysis.</i>	76
<i>The global assessment item as a merged impression of performance.</i>	77
<i>Generalizability analysis.</i>	79
RELATIONS WITH OTHER VARIABLES EVIDENCE.....	79
<i>Few junior resident participants.</i>	81
RESPONSE PROCESS EVIDENCE	81
<i>Data collection accuracy.</i>	82
<i>Effect of discipline.</i>	82
<i>Six-point rating scale.</i>	84

<i>Feasibility</i>	85
<i>Tool characteristics impacting the response process</i>	86
Entrustability-aligned rating scales.....	86
The place of rater training.....	87
CONSEQUENTIAL EVIDENCE.....	88
<i>Does the OCAT change formative feedback given in surgical clinic?</i>	88
<i>Formative feedback and trust</i>	91
<i>How do residents feel about receiving low scores on the OCAT?</i>	91
<i>Residents in difficulty and the OCAT</i>	93
LIMITATIONS OF STUDY.....	94
<i>Context limitations</i>	94
<i>Additional limitations</i>	95
STRENGTHS OF THE OCAT.....	95
FUTURE DIRECTIONS.....	96
<i>How are these OCATs going to be aggregated meaningfully?</i>	97
<i>How much rater training?</i>	97
<i>Paper or Electronic OCAT?</i>	98
CONCLUSION	99
ACKNOWLEDGEMENTS	100
REFERENCES	101
APPENDICES	117
APPENDIX A: THE CANMEDS ROLES.....	117
APPENDIX B: COMPARISON OF THE CANMEDS AND ACGME COMPETENCY FRAMEWORKS*.....	117
APPENDIX C: THE ZWISCH SCALE*.....	118
APPENDIX D: THE OTTAWA SURGICAL COMPETENCY OPERATING ROOM (O-SCORE) SCALE**.....	118
APPENDIX E: CBME CHALLENGES FOR SURGERY & POTENTIAL SOLUTIONS.....	119
APPENDIX F: SUMMARY OF VALIDITY EVIDENCE GATHERED.....	121
APPENDIX G: INITIAL REB APPROVAL LETTER.....	122
APPENDIX H: REB ADDENDUM APPROVAL LETTER.....	123
APPENDIX I: NATIONAL EXPERT SURVEY EMAIL REQUEST.....	124
APPENDIX J: ANCHOR SCALE COMPARISON: O-SCORE TO OCAT.....	125
APPENDIX K: RESIDENT RECRUITMENT EMAIL.....	126
APPENDIX L: STAFF SURGEON RECRUITMENT EMAIL.....	127
APPENDIX M: CONSENT FORM.....	128
APPENDIX N: PRE-PILOT RESIDENT SURVEY.....	132
APPENDIX O: OCAT FOR PILOT PROJECT (WITH VERTICAL LINE).....	133
APPENDIX P: CODING STRATEGY FOR SHORT ANSWER QUESTIONS.....	135
APPENDIX Q: FEEDBACK SURVEY AND FEEDBACK SESSION QUESTIONS.....	136
APPENDIX R: DISTRIBUTION OF RATING FREQUENCIES BY ITEM (1-13).....	137
APPENDIX S: SUMMARY OF FUTURE DIRECTIONS.....	150

Abstract

Purpose of Study: Workplace-based assessment is an opportunity for a learner to be assessed in their community of practice by an expert rater. The challenges and biases brought into this assessment relationship are complex. A shift towards Competency Based Medical Education in post-graduate residency education has triggered consideration of how to implement feasible assessment tools for the operating room, the in-patient ward, and the outpatient clinic.

Competent performance in outpatient clinic is vital to surgical practice, yet no assessment tool currently exists to assess daily performance of technical and nontechnical skills of surgery residents. This project describes the development of a competency-based assessment tool, the Ottawa Clinic Assessment Tool (OCAT).

Research Question: How does the OCAT demonstrate validity for measurement of surgical resident performance in clinic?

Method: A consensus group of experts was gathered to generate ideas reflective of a competent ‘generalist’ surgeon in clinic. An entrustability anchor scale was developed. A six-month pilot study of the OCAT was conducted in orthopedics, general surgery and obstetrics and gynecology with quantitative and qualitative evidence of validity collected. Two subsequent feedback sessions, and a survey for staff and residents evaluated the OCAT for clarity and utility.

Results: The OCAT was developed as a 13-item tool, with a global assessment item and 2 short answers questions. 44 staff surgeons completed 132 OCAT assessments of 79 residents.

Psychometric data was collected as evidence of internal structure validity and relations with other variables. Analysis of feedback indicated the rating scale was practical and useful for surgeons and residents.

Conclusions & Contribution to the Research Field: Surgical programs will require a daily clinic assessment tool to help define resident competency progression. Multiple sources of validity evidence collected in this pilot project demonstrate that the OCAT can measure resident clinic competency in a valid and feasible manner.

List of Figures

- 1 Seven-step Checklist for Developing a Good Assessment Instrument
- 2 Checklist for Consensus Group Technique
- 3 Steps of Consensus Groups
- 4 OCAT Entrustability Anchor Scale
- 5 Number of OCAT Forms Submitted by Individual Residents
- 6 Generalizability Equation

List of Tables

- 1 Missing Data Frequency
- 2 Glossary of Items on the OCAT
- 3 Consensus Group Resident Qualities Discussion
- 4 List of Items Following First Consensus Group
- 5 List of Items Following Second Consensus Group
- 6 Academic Affiliations of National Expert Survey
- 7 National Expert Survey Responses
- 8 Analysis of Responses from National Survey Compared to Consensus Groups
- 9 Responses from National Survey Not Appearing in Consensus Group List
- 10 Final List of Items and Descriptors
- 11 Number of Residents Consented for OCAT Study by Division
- 12 Pre-Pilot Survey Results - Resident Satisfaction with Current Feedback Rates
- 13 Free-Text Comments Received on Resident Pre-Survey
- 14 OCATs Submitted by Discipline
- 15 Item-Level Statistics for the Ottawa Clinic Assessment Tool (OCAT)
- 16 Inter-Item Correlation Matrix
- 17 Ottawa Clinic Assessment Tool Variance Components
- 18 Variance Components Explained
- 19 Descriptive Statistics of Items 12 & 13 (Subscale)
- 20 Technical Skills Subscale Q12 & Q13 Correlations
- 21 Global Assessment Item
- 22 Effect of Discipline Q1-Q11
- 23 Effect of Discipline Q12-Q13 Subscale
- 24 Effect of PGY Level Q1-Q11
- 25 Effect of PGY Level Q12-Q13 Subscale
- 26 Effect of Clinic Type Q1-Q11
- 27 Effect of Clinic Type Q12-Q13 Subscale
- 28 Comments Relating to Specific Elements of Clinic Done Well
- 29 Short Answer Questions – How Many Comments Correlated Directly to Items on the OCAT?

List of Acronyms & Abbreviations

ACGME	Accreditation Council for Graduate Medical Education
CanMEDS	Canadian Medical Education Directives for Specialists
CBD	Competence by Design
CBME	Competency Based Medical Education
DEC	Daily Evaluation Card
DOPS	Direct Observation of Practice Skills
EPA	Entrustable Professional Activity
EFPO	Educating Future Physicians for Ontario
FITERs	Final In-Training Evaluation Reports
Gen Surg	General Surgery
G-study	Generalizability Study
ITER	In-Training Evaluation Report
MD	Medical Doctor
Mini-CEX	Mini Clinical Evaluation Exercise
NGT	Nominal Group Technique
OCAT	Ottawa Clinical Assessment Tool
Obs/Gyn	Obstetrics & Gynecology
Ortho	Orthopedics
OSATS	Objective Structured Assessment of Technical Skills
OSCE	Objective Structured Clinical Exams
O-SCORE	The Ottawa Surgical Competency Operating Room Evaluation
PGY	Postgraduate Year
PROM	Patient-Reported Outcome Measures
RCPSC	Royal College of Physicians and Surgeons of Canada
REB	Research Ethics Board
STACER	Standardized Assessment of Clinical Evaluation Report
WBA	Workplace-Based Assessment

Section 1: Introduction to Thesis

Problem Statement

Competency based surgical education with its focus on frequent and regular assessment will require increased formative feedback from staff supervisors to guide residents through milestones towards completion of residency. One crucial aspect of residency that surgical residents must master is how to manage a surgical clinic. Unfortunately, formative feedback is not yet formally built into the busy surgical clinic environment, and a feasible in-training assessment tool for this situated environment (that promotes face-to-face feedback) is required. Currently there is no formative tool available that would allow one to assess competence in surgical clinic while demonstrating adequate validity evidence to support its use.

Study Objectives/Project Purpose

The goal of this study was to develop an assessment tool that would measure resident competence in surgical clinic and gather the validity evidence that would it use.

Thesis question. How does the OCAT demonstrate validity for measurement of surgical resident performance in clinic?

Sub-objectives.

1. To identify key elements of performance in the surgical clinic workplace in order to define the construct of interest for developing an assessment tool of surgical clinic competence.
2. To evaluate the feasibility of using the OCAT in surgical clinic.
3. To gather information on some of the challenges of workplace based assessment in surgical clinic. These consisted of elements of response process and consequential validity:
 - a. Respecting the expertise of surgeon supervisors and allowing them to use subjective impression in their assessment.
 - b. Giving formative feedback
 - c. Assigning low scores

Future Directions

This project involved the development of an assessment tool for a very specific area of surgical resident competence, namely the surgical clinic. Throughout this process many areas for further study in surgical competency-based assessment were exposed. The pilot project has successfully demonstrated both psychometric data to support tool validity, and qualitative commentary to support feasibility. Further recommendations for ongoing study would include using the feedback data gathered in this study to modify the tool and improve feasibility. The psychometric and qualitative performance of could then be tested through implementation in other surgical departments, helping to determine generalizability.

Future research should center on use of the OCAT for competency-based assessment as a formative WBA tool in practice. Given that study conditions will no longer be present (e.g. research assistant picking up the forms, dedicated study team to populate a database), elements of feasibility would need to be iteratively assessed. Consequential validity questions for future studies germinating from this pilot project include, but are not limited to: whether the tool should be paper-based or electronic, how often a residency program should expect their residents to have the OCAT completed, and how the OCAT will be used to contribute to the Final summative In-Training Evaluation Reports (FITERs) which program directors use to report to the Royal College of Physicians and Surgeons of Canada (RCPSC) as an indicator of readiness to sit the Royal College final exam.

Section 2: Background

Introduction

Post-graduate trainees, or resident physicians, finish medical school and complete the remainder of their training in the workplace. While treating patients they also handle multiple demands on their time and interact in an established social environment. As adult learners, resident physicians need to take responsibility for their own learning (Bennett, 2011). However, in many respects they are still novices and therefore require guidance from staff experts. This guidance comes through staff supervisor mentorship, as well as in the form of established objectives and regular assessment under the umbrella of Competency Based Medical Education (CBME).

A resident experiences graduated responsibility, receiving higher levels of independence as he or she progresses towards the ultimate goal of autonomous practice (Cogbill, Malangoni, Potts, & Valentine, 2014). How these progression decisions are made is a subject of debate that has led to a new method of training called ‘entrustment decision making’ (Ten Cate, 2005). Entrustment decision-making captures the performance of residents on their day-to-day tasks. One of the major challenges for medical educators today is tracking these assessment moments in a way that facilitates learning, but does not take away from patient care and the service residents provide to the healthcare system. Surgical educators are aware that a focus of the new CBME system is on documentation of these assessment opportunities to demonstrate progression of individual residents (Bhatti & Cummings, 2007).

Multiple assessment tools, with varying levels of feasibility for use in the surgical environment are available in the literature. Given that assessment can either be removed from the workplace environment (e.g. standardized testing, oral exams, or simulation testing), or performed as in-training assessment, program directors can pick and choose which tools they would like to use in their program of assessment (Schuwirth & van der Vleuten, 2011). Workplace-based assessment (WBA) tools, or in-training assessments, are a valuable resource within surgical CBME programs for tracking performance in a patient care context (Schuwirth & Ash, 2013; Ten Cate & Scheele, 2007).

The three major in-training surgical patient care contexts are: the operating room, the in-patient ward, and the outpatient clinic. Independent competence is necessary in all domains prior to graduation. Different skills are stressed in each one; for example technical skills are highlighted in the operating room and non-technical (or intrinsic skills such as professionalism

and communication) skills in the clinic. Thus far, tools for assessing non-technical skills in the outpatient clinic environment have been developed for observation of individual patient encounters only. There are currently no WBA tools available that surgeons can use to track their residents' performance over a day of surgical clinic. Therefore, the goal of this project was to develop a workplace-based surgical clinic assessment tool surgeons will actually use and to collect validity evidence supporting its quality.

Theoretical framework

The medical profession is one of the most respected in Canadian society, and practicing within it means significant responsibility. Training of new physicians is therefore an important task. Most applicants to medical school in Canada first attain a three to four year undergraduate degree. Therefore once a student graduates from four years of medical school to receive their Medical Doctorate (MD) degree they often have already completed eight years of post-secondary education. Following medical school there is a two to five year residency to complete (two for family medicine, four for internal medicine, and five for surgery). Residents have mixed responsibility: service to patients as a medical doctor (matched to their level of responsibility), plus the learner role expected of a resident physician-in-training. Balancing these two, often conflicting roles, of service and education is a defining feature of residency (Sonnadara et al., 2014).

The social nature of residency training contributes complexity as well. Social interactions with nurses, patients, and staff supervisors form the foundation for becoming a surgeon. From day one, residents learn by participating in an authentic environment, interacting with patients and team members, so that 'situated learning' is not an individual process but a collective one (Lave & Wenger, 1991). Learners rely on their co-residents, staff supervisors, and other members of the "Community of Practice" to demonstrate how to participate in the medical culture, not just how to complete technical tasks (Fuller, Hodkinson, Hodkinson, & Unwin, 2005). Learning to form relationships and going through the process of becoming a member of the professional community are vital components of residency (Rogoff & Lave, 1984; Wenger, 1998), making it much more than just a technical education.

Social constructivists consider the importance of community interactions and how they shape the learning process (Rogoff, 1995). Lave and Wenger (1991), who wrote on social constructivism, explain how social factors in the supervisor-learner relationship impact on responsibility-denoting decisions in a process they call "legitimate peripheral participation". In

this model, each member of the social community, from the most junior resident to the supervising surgeon, has a participatory role. Legitimate participation in the community of practice gives each individual purpose, a sense of belonging, as well as a right to ask for formative feedback (Fuller et al., 2005; Lave & Wenger, 1991). The postgraduate training equivalent of legitimate peripheral participation in action is graduated responsibility. Junior resident trainees work towards competence in their profession, receiving increasing levels of responsibility as they progress, and those in the center of expertise are responsible for helping them get there (Cogbill et al., 2014).

Medical education context: Competency based medical education.

To guide learners towards the center of expertise, medical educators need to know what it means to be a medical professional in their specialty, and consider how one gets there. The underlying question is what does it take to help junior residents become competent, independent medical professionals? The medical education community is answering this question with a curriculum model entitled Competency Based Medical Education (CBME). In short, CBME focuses on outcomes measurement and therefore requires sound assessment throughout residency (Holmboe, Sherbino, Long, Swing, & Frank, 2010; Ten Cate, 2013). Each specialty first needs to define the outcomes desired at final certification of their graduates, titled ‘competencies’ (Lurie, 2012). There is a fine distinction between the state of being *competent* and this idea of *competencies*, although they are integrally related.

To be competent (or possess competence) is the qualification of an individual to effectively and appropriately complete a task. The Royal College of Physicians and Surgeons of Canada (RCPSC) has recently defined competence within medical education as an array of physician abilities or performances that exist across multiple contexts considering case context and resident stage of training (Frank et al., 2010). Competence is multi-dimensional and changes with time, therefore requiring frequent re-assessment. As well, there are levels of competence, defying the binary calculation of competent versus not competent (Dreyfus & Dreyfus, 1980). Competence is also contextual, reflecting the relationship between a person’s abilities and how they complete the tasks in a particular situation. The goal of CBME is for residents to reach a level of *independent* competence measured as a series of achievable outcomes over time.

In contrast, a competency is a desired characteristic or function of a fully mature physician. The RCPSC defines a competency as an observable ability of a health professional related to a specific activity, integrating knowledge, skills, values, and attitudes (Frank et al., 2010). The

‘observable’ portion of this definition is key, as it implies these are measurable traits that can be assembled to facilitate progressive development of a trainee. To ease CBME curricula being instituted in their specialty, many groups have begun defining competencies specific to their learning objectives modeling frameworks such as the Canadian Medical Education Directives for Specialists (CanMEDS) roles (Bancroft et al., 2008; Korte et al., 2013; Lee, Oetting, Beaver, & Carter, 2009).

CanMEDS Competencies. In Ontario in the early 1990s, the Educating Future Physicians for Ontario (EFPO) project was launched to determine what Ontarians expected of their future physicians and how the educational system should adjust to respond (Maudsley, Wilson, & Neufeld, 2000). The RCPSC approved the project and it was appropriated under their leadership as the CanMEDS project, promoting a discussion of traits the public would most like to see displayed in their own physician. Seven of these domains were proposed for ranking (advocate, manager, scholar, communicator, medical expert, collaborator, professional). However, the response was strongly in favor of all the roles being equally important, and so they were not ranked. Thus the equally weighted “flower” of CanMEDS was developed (see Appendix A) ushering in an era of CBME structured around these competencies (Chou, Cole, McLaughlin, & Lockyer, 2008). It is interesting to note the breadth of personal competencies contained within the CanMEDS roles with the traditional view of the physician as a medical expert nearly overshadowed by other non-technical roles (Arora et al., 2009).

The CanMEDS competencies are being revised beginning in 2015 to incorporate the full spectra of training – from early training through to continuing professional development. This new ‘competence continuum’, titled Competence by Design (CBD), de-emphasizes time-based elements of training and stresses lifelong learning principles. Besides aligning the CanMEDS roles to CBME with the introduction of milestones and entrustable professional activities (EPAs), a major goal of CBD is to honor requests from frontline educators to make the roles easier to teach and assess (Royal College Website, 2015).

The Accreditation Council for Graduate Medical Education (ACGME), the non-profit organization that reviews and accredits graduate medical education programs in the United States, has created a parallel system over the last few decades. This system contains six very similar competencies: patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice (Rushton, Hicks, & Carraccio, 2010). See Appendix B for a comparison of the

CanMEDS and ACGME roles (Hodges, 2010), demonstrating how the main differences exist between terminologies while the spirit of the roles is maintained across the two systems.

Entrustable professional activities and milestones. Despite the CanMEDS competencies being defined as abilities, not actions, there is still a tendency to break down the competencies into checklists of tasks to be executed (Glass, 2014). The tendency to write checklists has fueled a counter-CBME argument, which insists deconstructing such complex ideals washes out their original intent (Talbot, 2004). To address this tendency, ten Cate (2005) introduced a perspective on competencies known as the Entrustable Professional Activity (EPA). His paper argues that competencies (e.g. the CanMEDS competencies) should be acknowledged as the overall goal to be achieved, but in the day-to-day work environment, activities are the elements of professional work. An EPA, as defined by ten Cate, is a core unit of professional work identified as a task or responsibility to be entrusted to a trainee once sufficient competence has been reached (ten Cate, 2005). EPAs address a professional task, with a recognizable output, that can be observed by an expert (Meade et al., 2012). In this way, EPAs make the idea of progression decisions towards competence practical for frontline educators by giving benchmarks independent of the resident's level of training.

A term often confused with EPAs is milestones. In the updated CanMEDS 2015 framework, the RCPSC has specified the key difference for the Canadian postgraduate system is that EPAs are the tasks of a profession, and milestones are achievable aspects of that task that an individual must master on the road to completing EPAs (Royal College Website, 2015). Writing discipline-specific milestones and their end-goal EPAs has been accomplished in the United States (Rushton et al., 2010; Santen et al., 2013; Cogbill et al., 2014), and by most specialties in Canada. As this is a recent development, many programs are still working on bringing them into common assessment usage within their local environment.

Entrustability Decision Making and Construct Aligned Scales. After defining milestones and EPAs, a training program needs to develop targeted and feasible assessment tools so frontline clinical educators can assess how each resident is progressing towards completing them. In order for assessment tools to be accepted in clinical practice, it is crucial frontline educators feel an assessment tool captures their true appraisal of a resident. Medical education has seen an evolution in assessment tool design, moving away from checklists of objective criteria (Brooks, 2009; Taber et al., 2010) towards formats that accept value in capturing the subjectivity of expert raters (Eva & Hodges, 2012). One of the arguments against listing

milestones, and ticking them off one-by-one as they are completed, is that there is a risk of missing subtle resident qualities when attempting to be too objective (Jones & Phillips, 2014; Ginsburg, McIlroy, Oulanoya, Eva, & Regehr, 2010). Also, depending on how milestones are written they may represent general abilities, which can only be assessed across rotations and clinical encounters. Workplace based assessment tools used by only one staff physician at a particular point in time would therefore not help to assess milestone attainment. Evidently then, just having defined CanMEDS roles, milestones and EPAs will not automatically create viable assessment tools.

Making milestones and EPAs ‘assessable’ in clinical practice has been a work in progress. It has been shown that clinical supervisors can struggle to turn their observations of resident clinical performance into checkmarks or ratings on a traditional assessment form. Yeates, O’Neill, Mann, and Eva (2013) refer to this as “information integration” difficulty (Yeates et al., 2013), saying that assessment decisions are often not easily converted into numerical or categorical format. It is possible that this tension is even more real when assessing non-technical (non-procedural) skills such as communication (as is the goal with the assessment tool developed in this study). In the communication-based field of social work, Bogo et al. (2006) identified a potential explanation for why assessment tools based solely on objective criteria, and not built with raters in mind, fail to be accepted as capturing valid assessments. In Bogo’s study, supervisors were asked to vocalize their ratings of students’ performance and a study team member mapped them onto an assessment form. But in fact, they did so in such a complex way, they were unable to map back responses to the competency domains used on traditional assessment forms.

Ten Cate and Scheele (2007) noticed the phenomenon of information integrating difficulty in medical education assessment and proposed a new way of thinking about assessment decisions (rather than by competencies or milestones). Their proposed way of thinking, titled entrustment, was found to be the way clinician supervisors across specialties think about their interactions with residents. They found that staff physicians working with learners routinely ask themselves, “Can I leave this resident alone?” or can “I entrust this task to this resident?”. The construct of interest for the assessment tool proposed in this thesis project is ‘surgical resident clinic competence over a day of clinic’. From the beginning, a goal was to write the tool in such a way that allowed staff physicians to use this natural entrustment questioning when completing their assessment.

Construct aligned scales. Having found that entrustment is a common language spoken by staff supervisors, Crossley (2011) took the idea one step further in a landmark study that simply changed the wording on assessment scales to be aligned with this cognitive construct. His team re-wrote assessment tools in language that better reflected how faculty supervisors conceptualize the practice of their trainees (Crossley, Johnson, Booth, & Wade, 2011). ‘Construct alignment’ as they called it, gives voice to the more ‘subjective’ impressions of the experts who seem able to make highly reliable judgments when they are asked to in the right way (Crossley et al., 2011; Ginsberg, Gold, Cavalcanti, Kurabi, & McDonald-Blumer, 2011). Construct-aligned anchor scales provide evaluators with a standardized assessment measure that is structured around the way evaluators already make day-to-day decisions. It may not be that raters disagree about the performance observed, but more that they translate it into a numerical rating scale unreliably (Hansen, Rosenberg, & Lane, 2013). Rather than asking evaluators to make assessments against abstract scales, such as cohort ranking or post-graduate year, Crossley and his colleagues proposed a new type of scale based on construct alignment.

Entrustability scales. Entrustability scales are an example of a construct-aligned anchor scale, in that they acknowledge how a staff supervisor is thinking about their everyday teaching experiences. Examples of the performance of this type of scale are occurring in the assessment literature recently. The Zwisch scale (Appendix C), a behaviorally anchored ordinal scale (1-4) used to grade the degree of guidance necessary during a technical procedure, is a great example of an entrustment anchor scale (George et al., 2014; DaRosa et al., 2013). At the lowest end of the scale is “show and tell” where each step is outlined for the trainee, and at the top, “supervision only” at which the supervisor’s presence is warranted only to ensure patient safety. Another example of an entrustability scale is that used in the Ottawa Surgical Competency Operating Room (O-SCORE), as seen in Appendix D (Gofton, Dudek, Wood, Balaa, & Hamstra, 2012). In both of these scales, raters are still asked to choose a number but the process has more meaning to them as their real-world judgments are reflected.

Entrustability scales also draw attention to a trainee's readiness for independent practice (the goal of CBME), rather than how they relate to their peers. A resident has years to strive for independence in their EPAs, and should be able to both track their progression and identify deficiencies using an entrustability scale (Ten Cate, 2005). Medical educators are aware that a focus of the new CBME system is on documentation of graduated responsibility, and assessment tools that include entrustability decision making show promise for making this process simpler.

Surgical Education

Brief introduction to surgical education. CBME has been quickly introduced from theory to practice, creating an understanding disconnect for frontline surgical educators who may be unfamiliar with new terminology like competencies and EPAs. It is not merely terminology that has outpaced the educational training of most academic surgeons, but also the latest updates in feedback delivery approaches and assessment methodologies (Sanfey, 2014; Epstein, 2007). Combined with the pressures of fiscal reform, technological development, and increasing accountability, rapid educational change has the potential to alienate crucial elements of the postgraduate Community of Practice (Mellinger, Damewood, & Morris, 2015). Moving forward, efforts will need to be initiated by surgical educators who understand the goals and constructs of CBME, in order to create a system accessible to frontline educators who are responsible for resident assessment. Those developing new assessment tools, especially tools that are intended to be used in the workplace, need to keep this in mind.

Surgical assessment context. Most surgical residency training programs in Canada are five years long, with the potential for additional years of fellowship training. The three major in-training surgical contexts are the operating room, the in-patient ward, and the outpatient clinic, with independent competence necessary in all domains prior to graduation. Trainees encounter a variety of supervisors and surgical practice styles during this period (Gillespie, 2005). Given that residency is one of the most crucial time points in the professional training of surgeons, ensuring that newly trained surgeons have met competency standards prior to independent practice stands as the primary objective of this period.

But what is surgical competence? Is it solely the ability to operate safely or is the profession more than that? One of the highest complements a surgical resident can pay to their supervisor is that they would trust their family members to their care. This trust is often based on more than slick technical ability; it includes a combination of bedside manner and hard-to-identify factors like respect. Michael Kane (1992) acknowledged this when he defined clinical competence as a combination of knowledge, skills, and judgment rolled into a profession (p.167).

Training a surgeon is more than teaching surgical technical skills; surgical competence is the ability to apply professional knowledge, skills, and attitudes to novel situations inside the operating room and out (Bhatti & Cummings, 2007). Surgical competence encompasses the

cognitive skills to solve clinical problems (clinical reasoning), the ability to operate safely, and the psychosocial skills to communicate with patients, families, and the interdisciplinary team. As Van Eaton et al. (2011) comment, surgery is a lot more than just cutting and tying. It takes more than just putting a resident in a hospital with patients for five years to create the knowledge and judgment needed to be a competent surgeon (Van Eaton et al., 2011, p.469).

This range of skills and abilities are both modeled implicitly by the Community of Practice and taught explicitly, but regardless, recent changes to the surgical training environment will require explicit assessment of all surgical competencies. Adding an increased assessment load to surgical training programs will be difficult since other external pressures on residency programs are also increasing. These additional pressures include: the reality that residents are actually working fewer hours due to work hour restrictions (Lewis & Klingensmith, 2012), increasing intolerance of error by the public and the hospital system, ever-present financial constraints, challenging technological advances for residents to keep up with (Grantcharov & Reznick, 2009), and an appropriate focus on patient safety which can drive hospital policy against the resident's learning experience (Sanfey, 2014). In addition, there are numerous reports of surgery residents feeling their training is inadequate, with some residents leaving residency not feeling confident they can independently (Maxwell et al., 2010; Fitzgerald, Marron, & Giddings, 2011). Therefore residency is being compressed into fewer hours and less time in the operating room (Purcell & Tarpley, 2009), and surgical training programs must somehow become more efficient, maintain or increase effectiveness, and learn to create more space for documenting learning and assessment moments.

To cope with these pressures, the RCPSC in Canada has made plans to implement CBME programming and assessment in surgical residency programs. For most program directors this means that prior to accreditation time, they need to be able to articulate how their program is (1) teaching, (2) assessing, and (3) documenting the progression and potential remediation of each individual resident. In terms of CBME implementation, there are residency programs that have piloted completely non-time-based approaches. The orthopedic residency program at the University of Toronto just graduated the first of their CBME classes that was completely non-time based (Alman, Ferguson, Kraemer, Nousianien, & Reznick, 2013). Preliminary, yet unpublished, results of this program show the residents enrolled in the CBME stream are meeting competency milestones and in fact are reaching it at a faster rate than their mainstream colleagues. Although the speed of resident progress is not, and should not, be the focus of CBME curricula, the Toronto experience has increased interest in the outcomes focus of CBME.

Toronto's involvement has also highlighted challenges of CBME at the programmatic level specific to the surgical environment. See Appendix E for a table illustrating these challenges as this discussion is beyond the scope of the current paper. Moving forward, key considerations for *all* surgical programs implementing a CBME curriculum will be the presence of solid assessment programs and workplace-based assessment (WBA) methods that capture the true perspectives of surgeon raters. This thesis project attempts to create such a WBA tool for use in a CBME surgical environment, incorporating entrustment decision-making and graduated responsibility theory.

Workplace-based assessment in surgery. As most surgical education occurs concurrently in the patient care workplace, it would be ideal to be able to assess in the workplace as well. Unobtrusive, 'real-world' assessments seem to be a better way of documenting "normal" performance (Kopelow et al., 1992; Sackett, Zedeck, & Fogli, 1988), and WBA is thought to be the optimal method for assessing professional competence (Epstein & Hundert, 2002). Compared to other methods of assessment which capture a resident *knowing how* to do a task (written and oral exams) or *showing how* to do a task (Objective Structured Clinical Exam or OSCE), the purpose of WBA is to capture a resident *doing* a task in their every day environment. Although all levels of assessment have their place, Miller believed what a student *does* to be the greatest indication of what they have learned and placed it at the top of his well-known assessment pyramid (Miller, 1990). Ideally in residency we want to assess what residents actually do, not what they potentially could do on an examination.

WBA evolved out of early work done by Lave and Wenger (1998) critiquing assessment assumptions that are largely still with us. In modern education, learning has been considered an individual process, and assessment has been largely separated from the rest of workplace activities for the sake of standardization and objectivity. Wenger states the dilemma this way,

To assess learning, we [educators] use tests with which the students struggle in one-on-one combat, where knowledge must be demonstrated *out of context*, and where collaborating is considered cheating (Wenger, 1998, p.3).

If the desire is to bring assessment back into the training environment, a good starting point is to question why WBA has not been taken up widely in surgery so far. The most likely reason is lack of time and resources (Sonnadara et al., 2013). The three major in-training surgical contexts are the operating room, the in-patient ward, and the outpatient clinic, with independent competence necessary in all domains prior to graduation. The focus in all three is

on timely, safe patient care, with varying priority levels placed on resident education. Depending on time available, the staff surgeon may feel that adding a WBA takes away from time spent with patients (Jones, van Kessel, Swisher, Beckstead, & Edwards, 2014). Conflicting priorities mean that the majority of surgeons working in academic centers would like to give more complete, high quality feedback to their residents, but find the usability of existing assessment instruments poor (Grantcharov & Reznick, 2009, p.106). Compensation for time spent teaching and assessing is variable (Goitein, 2015), and resources designated for training staff how to properly use assessment tools are often limited (Dath & Iobst, 2010). Therefore time constraints cause decreased numbers of assessments completed, as well as decreased time for rater reflection and feedback.

A second reason for low uptake of WBA in surgery is an over-reliance on using one assessment method. Most surgical programs use solely end-of-rotation In-Training Evaluation Reports (ITERs) at this time (Sherbino, Kulasegaram, Worster, & Norman, 2009). One faculty member, who may or may not be supervising the trainee, completes an ITER at the end of a rotation (usually a rotation supervisor, not necessarily the program director). Therefore feedback given is usually delayed, and rarely based on concrete examples. Ideally, instead of relying on one method, a program director should design an assessment program combining multiple tools into an overall picture of how the resident is progressing towards competency (Bok et al., 2013).

Finally, infrequent surgical WBA results from concerns over subjectivity due to changing context and the social biases of raters (Kogan, Holmboe, & Hauer, 2009). In the workplace-environment it is nearly impossible to control all variables and produce a similar assessment context between encounters. This introduces significant subjectivity and variance. In the past, subjectivity concerned medical educators enough to cause them to prevent WBA assessment from occurring. Therefore if CBME educators are going to use WBA, it is important for them to understand the sources of bias that have been so concerning.

Biases in surgical workplace-based assessment. WBA clearly depends on raters (expert or not), who are asked to judge performance. In a recent review of in-training assessment, van der Vleuten and Verhoeven (2013) note that the value of assessment instruments depends more on the users (raters), than on the instruments themselves (Van der Vleuten & Verhoeven, 2013, p.458). The social dynamics of the system affect rater judgment and decision-making processes. Human judgment is an indispensable necessity in a program of assessment (Schuwirth & van der Vleuten, 2011) and raters can be viewed as active information processors who both

interpret and construct their personal impression of the assessment context (Govaerts, Van der Vleuten, Schuwirth, & Muijtjens, 2007, p.252). Therefore assessment in a clinical environment can be a decision that is prone to bias.

These rater biases have been well described (Yeates et al., 2013; Kogan, Conforti, Iobst, & Holmboe, 2014; Gingerich, Regehr, & Eva, 2011; Williams, Klamen, & Mcgaghie, 2003) and include such effects as: ‘failure to fail’, leniency bias, and the polarization of raters. Medical education has long known that there is a “failure to fail” (Dudek, Marks, & Regehr, 2005) and supervisors struggle to define lapses in adequate performance (Cleland, Knight, Rees, Tracey, & Bond, 2008). The social nature of the supervisor-trainee relationship discussed above can introduce complex emotional implications potentially contributing leniency into the assessment encounter. For example, it can be more difficult for supervisors to assess a trainee in what may be perceived as a harsh manner once a comfortable social connection is established.

This leniency bias is compounded by the tendency of supervisors to communicate pleasant messages (the ‘Mum effect’) and to avoid unpleasant ones (Tesser & Rosen, 1975). Yeates et al. (2013) describe the lack of productive, constructive feedback language used effectively in the title of their paper entitled, “You’re certainly relatively competent” (Yeates et al., 2013). This type of assessment communication is all too common.

Another common rater bias is seen in the tendency of raters to adopt a rating style – whether lenient or stringent – earning them the title ‘dove’ or ‘hawk’ respectively (McManus, Thompson, & Mollon, 2006). The danger here is that due to this bias, a resident who is not meeting expectations but is rated only by a ‘dove’ may not be identified as such and pass through to the final years without receiving any remediation. There are statistical ways of accounting for these distinct rater differences (Baker, 2011), however ensuring multiple raters capture a variety of encounters over time means the rating styles of individual supervisors become averaged out (VanderVleuten & Schuwirth, 2005).

Other social factors, besides natural grading tendencies, likely compound a supervisor’s tendency to grade leniently. From the surgeons’ perspective, there are increasing challenges associated with giving a negative review. A negative assessment will likely mean increased paperwork, potential legal action by the student, and requests from administration for ‘more evidence’ (Dudek et al., 2005). Understandably, this may cause raters to just want to avoid confrontation.

Even though raters and trainees often know each other and have a friendly social relationship, surgeon supervisors still frequently depend on first impressions of their residents’

true abilities due to their sporadic interaction. Paying specific attention to how first impressions guide our assessments is an emerging field in medical education (Wood, 2014). It has been known for some time that assessors' first impressions of students may be more important to how they rate than the actual performance (of case presentations) (Wigton, 1980). As well, trainees are known to act differently when aware they are being observed (aka the Hawthorne effect) (Rees & Knight, 2007). For these reasons, it is important to find ways to sample in both obtrusive and non-obtrusive ways as often as possible (Kopelow et al., 1992; Sackett et al., 1988).

Although the influences of contextual, social, and cognitive biases on WBA are much more complicated than described above, an extended discussion of this topic is beyond the scope of this thesis. Surgical educators and program directors attempting to increase the use of WBA should be aware of the both the challenges and limitations of workplace based assessment while being able to develop an assessment program that controls or anticipates their influence. Some potential ways to strengthen WBA have been considered in this project and will be outlined in the discussion.

Workplace-based assessment options in the literature. Multiple assessment tools, with varying levels of feasibility for use in the surgical environment, are available in the literature. Determining the appropriate number of assessment tools for a program will need to balance the benefit of multiple sources of assessment opportunity, with the learning curve and rater training necessary when a new tool is introduced. There are numerous workplace-based tools available in the literature, however very few of them are targeted to the surgical context (Cogbill et al., 2014).

Assessment opportunities within surgical residency can be distributed between training environments (i.e. not all assessment opportunities are in the operating room, or the simulation lab). Frequently the focus is on operative skills, but surgical residents must also master running a surgical clinic independently. Clinic is where patients and their families first meet the surgeon to decide if surgery is needed, and where the surgeon receives informed consent to proceed if it is needed. Following an operation, most surgeons will see patients in follow-up clinic to ensure proper wound healing and answer any questions patients may have. Complex non-technical skills are required in order for high quality care to be provided in clinic.

Previous WBA tools for use in clinic are based on direct observation of the resident's interaction with a patient (Bandiera & Lendrum, 2008; Wilson et al., 2014). As surgical residents receive more graduated responsibility however, they are rarely directly observed in the

clinic environment. Instead, they spend the day seeing patients concurrently with the staff and asking for help or advice when needed. The staff surgeon therefore develops an idea of their ability to manage the clinic across an entire day of clinic, and over multiple patient encounters. Current clinic assessment tools cannot account for this (Kogan et al., 2009).

For example, the Mini-Clinical Evaluation Exercise (mini-CEX) is a frequently cited WBA tool developed for the clinic environment. This is the best-known single performance tool (Norcini, Blank, Duffy, & Fortna, 2003), with more than 20 studies demonstrating its reliability and validity properties (Kogan & Holmboe, 2013). The mini-CEX is used to assess a specific encounter with one patient as an explicit opportunity (Norcini, Blank, Arnold, & Kimball, 1995). Originally it was developed for feedback purposes (Norcini et al., 2003), and only later have programs start using it in a more summative role. A strength of the mini-CEX is that it does not require standardized patients and therefore is cheaper and more broadly applicable and can be flexibly applied in a number of clinical settings. It has also been modified specifically for the surgical environment (still based on a single patient encounter) as the CAMEO (Clinical Assessment and Management Evaluation – Outpatient), which shows similar psychometric properties to the mini-CEX itself (Wilson et al., 2014). However, this form is often not a good option for surgical clinic due to its reliance on direct observation, and how it only tracks resident performance with one patient. Crucial skills such as time management of a hectic clinic, and prioritizing complex versus routine patient cases are not considered in the mini-CEX.

A newer tool that resembles the mini-CEX is the Standardized Assessment of Clinical Evaluation Report (STACER) (Wright & Langille, 2011). This tool makes use of simulated, well-scripted scenarios designed around very specific competencies. Although it takes advantage of real patients as opposed to an actor playing a scripted role, the concern still exists that we are assessing “exam day” performance, not real world performance.

A variety of additional of scoring systems have been developed to assess performance in the clinical environment (not necessarily the outpatient clinic), including: the Daily Encounter Card (DEC), ITER, Case-Based Discussion, and 360-degree assessments (multisource feedback) (van der Vleuten & Verhoeven, 2013). DEC's are brief comment cards deployed after a defined experience and their design varies widely to include numerical ratings, qualitative comments, or a combination of the two (Brennan & Norman, 1997). Regular, documented assessment is a strength of using DEC's (Bandiera & Lendrum, 2008). Emergency medicine has developed a daily assessment card to be used across an entire shift (Sherbino, 2009). However, there is a concern with these CanMEDS daily encounter cards that supervisors do not find them self-

explanatory (due to lack of construct-alignment), causing low reliability and inability to discriminate between the CanMEDS roles (Sherbino, Kulasegaram, Worster, & Norman, 2013).

Other single encounter clinic WBA tools include: the Direct Observation of Practice Skills or DOPS (Wilkinson et al., 2008), chart stimulated recall (Jennett, Scott, & Atkinson, 1995), the Case-Based Discussion (Davies, Archer, Southgate, & Norcini, 2009), and the Professional mini-evaluation exercise or P-MEX (Cruess, McIlroy, Cruess, Ginsburg, & Steinert, 2006). These tools target either individual CanMEDS roles (eg. professional) and/or involve a large input of financial and administrative resources.

The Objective Structured Clinical Examination (OSCE) is an example of an assessment tool that has been used to assess the medical expert role in a more formal, standardized fashion. The goal is to re-create a realistic patient care scenario in a test environment, so that test-takers will feel like they are interacting with an actual patient, not a standardized one. Put in the terms of Miller's pyramid, the OSCE requires the student to *show how* they would complete the task in the real world outside of the workplace (Miller, 1990). The OSCE has now been adapted to evaluate non-technical skills in the surgical environment, albeit with a great investment of time and money (Davis & Lee, 2011). The main concern with this type of assessment is that it involves a simulated patient scenario, as opposed to a real clinical setting. Although useful, we need tools that do not assess exam performance but rather "everyday" performance to complement the more standardized exams in residency.

Although single-patient encounter assessments remain important, it is often more feasible in workplace assessment to include care provided to populations (or groups) of patients instead of individuals (Kogan & Holmboe, 2013). Portfolios are one example of an aggregate method that have been shown to increase the reflective abilities of medical students (Buckley et al., 2009) on their care of multiple patients. Patient-reported outcome measures (PROMs), a WBA method targeting meaningful outcomes for patients, are being adopted in the United Kingdom as well (Fitzpatrick et al., 2013). From the perspective of patient-oriented care, this may be an important source of assessment information residency programs can tap in the future. However, at this time Canadian surgical residents work in a self-regulating profession and therefore need to be assessed by their supervising surgeons and PROMs could therefore be at most a useful assessment program adjunct.

The most commonly used aggregate WBA tool in Canada is the ITER (Sherbino et al., 2013). These are used as mid- and end-of-rotation assessments upon which decisions are made regarding residency training (i.e. pass or fail of a rotation) (Turnbull & Van Barneveld, 2002).

Recently, there has been additional use of 360-degree assessments from interdisciplinary staff in order to gain a summative assessment of how the resident is performing in the domains of practice that are difficult for the supervising surgeon to assess (i.e. how well the resident interacts with nursing staff when on call) (Lockyer, 2013). The concern with both of these summative assessments is that often the supervisor lacks enough detail to complete the form appropriately, due to lack of direct observation. As well, there are concerns that retrospective bias contributes significantly since supervisors are often filling out ITERs a month or more after the resident completes the rotation (Williams et al., 2003). This causes poor reliability (Sidhu, Grober, Musselman, & Reznick, 2004) and residents can more easily shrug off any constructive feedback offered as unfounded. It has been shown that although there is great value in narrative comments on ITER forms, the opportunity to complete them is rarely taken (Dudek, Marks, Wood, & Lee, 2008). This is likely influenced by time lapse to completion date and completion of the form by someone other than the person who worked most closely to that resident. More tools that assess performance at the point of occurrence, avoiding retrospective bias, are needed to improve the quality of summative assessments and feedback.

Summary of workplace based assessment options in the literature. To summarize, WBA tools are a valuable resource within surgical CBME programs for tracking graduated responsibility in both the operating room and the surgical clinic (Ten Cate & Scheele, 2007; Schuwirth & Ash, 2013). The clinic environment is a vital area of surgical practice and performance, but most surgical programs do not have a targeted assessment tool for daily use, as most existing tools are based on simulated or observed individual patient encounters. The most commonly used tools are the mini-CEX and the ITER, the former infrequently used due to infrequent direct observation, and the latter suffering from significant retrospective bias. More standardized, resource-intensive tools such as the OSCE and STACER may have their place in an assessment program, but are not easy to incorporate into the current surgical clinic environment.

Validity Evidence

Collecting validity theory evidence. To document that the results of an assessment tool reflect its intended meaning, that is that the results are interpreted correctly, a necessary step in the development of that tool is to collect evidence that supports its intended use (Downing, 2003; Messick 1988; Cook & Beckman, 2006). This validity evidence comes from multiple

sources and has been divided into five categories: content, response process, internal structure, relations to other variables, and consequences (American Educational Research Association, APA, & NCME, 2014). A recent review of directly observed workplace based assessment tools states that minimal validity evidence has been gathered for most WBA tools and this should change (Kogan et al., 2009). Gathering multiple sources of validity evidence for the surgical clinic assessment tool in development during this project was vital to making an argument that it should be used in a competency-based assessment program in the future.

Content evidence. The first component of construct validity evidence collected is called content evidence. The aim of content validity is to make the items of the test purely represent the construct, and nothing else. A large part of tool development is to gather support for the relevance of the component items included, and to hear from stakeholders whether any of the construct is over- or under-represented. Content evidence is gathered throughout the development process, from item writing to expert revision and monitoring.

As a first step in determination of content, some consideration should be given of the purpose of the test. Will it be formative, summative, or potentially attempt to fulfill both designations (Schuwirth & VanderVleuten, 2011)? This decision potentially affects the way items are written (Hamstra, 2012). A summative assessment tool is written in such a way that a pass/fail designation can be made for purposes of grading. In contrast, a formative tool is built mainly to educate raters and trainees in areas that need remediation and further teaching focus.

Internal structure evidence. In this project, reliability data was considered evidence of internal structure as described by Downing (2004). There are other ways of discussing internal properties that this project's methodology does not explore (item response theory, Rasch modeling). Reliability means searching for consistency in a measure over time, over raters, or over encounters, and is considered a component of validity here. Internal structure validity is strongest when many assessments are collected.

Relations with other variables evidence. The goal of relations to other variables evidence is to compare a tool's results to the results of other similar tests (hopefully a gold standard, but not necessarily). There is also the expectation that results of a new tool would correlate with other similar tests. If there were no test of the same construct available in the literature, the expectation would be to compare results to those expected with application of common sense. Seeing progression in a group that you would expect to see progression in is

evidence of this type of validity.

Response process evidence. The fourth category of evidence is termed response process. How respondents interpret, elaborate, and reflect on the content of a test and whether this behavior corresponds to the original purpose is response process (Downing, 2003). Checking for response process validity helps pick up writer biases, since test taker interpretation may highlight differences in understanding.

Some components of response process validity exist for quality assurance. Attentiveness to data security measures and adequate scoring fit in this category. When data entry is inaccurate, responses are inaccurately portrayed and therefore it becomes a response process issue. Test format and logistics can also affect validity. For example, does having the assessment tool available in paper format versus electronic format affect the way the raters think about the items?

Although not formally considered to contribute to validity evidence, the feasibility of a tool is also a response process validity issue. A significant amount of time should be invested in ensuring feasibility, or the tool will not complete its intended purpose. For example, if a five-page WBA tool would be unlikely to be used outside of a study environment, then one would want to learn this early so it could be modified.

Consequential evidence. The fifth measure of validity evidence is consequential validity. Assessment tools have intentional and unintentional impact on learners and other stakeholders when administered. Evaluating the intended, unintended, acute, and chronic consequences of the tool is the most debated aspect of validity (Downing, 2003). Compared to consequential evidence, the other sources of validity evidence are all related to the scores a tool produces, rather than the tool itself. Consequential validity evidence involves assessing the impact of the tool on all the stakeholders, and how they will benefit or suffer as a result of it. Although debated and seldom reported (Beckman, Cook, & Mandrekar, 2005), consequential evidence will be discussed here because in the case of a new tool it can help with advance planning and harm reduction strategies. If consequential issues are considered from the beginning a plan to combat negative consequences of the tool can be put in place (e.g. remediation plans for struggling residents).

Summary of Background

Surgical residents learn to run an outpatient clinic and mature their non-technical skills in a

social workplace environment within a Community of Practice. Staff surgeons are responsible for guiding them from the zone of peripheral participation where they begin as newly graduated medical doctors, towards the central goal of competent independent surgeon. Guideposts along the road to competence are marked by regular assessments and the entrustment of tasks by multiple supervisors. Workplace based assessment provides the opportunity for supervisors to comment in-the-moment with directed feedback, but users need to be aware of its potential for bias. There has been some discussion of how WBA fits into competency based curricula, but currently there are few feasible assessment tool options to help staff surgeons keep track of their residents' daily progress in clinic, or to facilitate daily feedback to these same residents. WBA tools in existence focus mainly on direct observation of patient encounters, which rarely occur in a surgical clinic, or are completed too far after the fact to hold any significant meaning for resident learning. The remainder of this paper describes the development of and gathering of evidence for a construct-aligned WBA tool, specifically designed to measure resident ability to manage multiple patients over a day of surgical clinic. The goal of this project is to develop an assessment tool that surgeons will actually use in clinic to facilitate ongoing formative feedback, as well as collect validity evidence supporting the quality resulting from the tool's scores.

Section 3: Methodology

Summary - Introduction

To provide structure during the tool development process, and ensure all facets of validity evidence were gathered, the seven steps of test development suggested by Hamstra (2012) were used as a model (see Figure 1). These steps will be highlighted sequentially as they are addressed.

Figure 1: Seven-step Checklist for Developing a Good Assessment Instrument

1. Determine the purpose of your assessment
 - A. Formative, summative (standard setting/criteria) research
 - B. Knowledge, skills, attitudes (e.g., performance, teamwork, anxiety)
2. Content validity—identify main construct of interest and stakeholders
3. Review with content experts—focus group
 - A. Representative sample: different institutions and disciplines
 - B. Thematic saturation, address political issues
4. Item writing/development
5. If necessary, train the raters
6. Pilot test the instrument (representative sample) for validity
 - A. Feasibility check—length, clarity, cost
 - B. If necessary, go back to Step 4 (modify items and pilot test again)
7. Implement modified test—measure reliability, validity based on larger sample
 - A. Assess construct validity

To summarize, first it was important to define the construct the assessment tool would be measuring, namely surgical resident clinic competence over a day of clinic. When a literature review demonstrated no clear definition of surgical resident clinic competence and no assessment tools available for our intended purpose (see *Workplace Based Assessment Options in the Literature*), stakeholders were gathered. Two consensus groups of experts (residents, program directors, and staff surgeons) were held and a national survey of surgical educators was distributed to gather content for the test items. A short pre-pilot study checked for gross interpretation issues. The tool was named the Ottawa Clinic Assessment Tool, or the OCAT.

Following this, a six-month pilot study tested the psychometric and feasibility properties of the tool in three surgical divisions at the Ottawa Hospital. The principal investigator gathered

informal feedback throughout the pilot study from users of the tool and two team members entered data into a database on an ongoing basis. Finally two formal feedback sessions were held after the study closed, to assess stakeholder opinion of feasibility of the tool. A post-pilot project survey was also sent out to all study participants to gather additional feedback on the feasibility of the tool.

A summary of validity evidence collected throughout the process of tool development can be seen in Appendix F. Note that response process evidence is a strength of the tool, as stakeholders were involved throughout the entire study, giving informal and formal feedback. Content evidence was gathered both locally and nationally. Internal structure evidence was gathered from multiple sources including brief rater training sessions. Relations to other variables evidence is based on correlation with advancing postgraduate (PGY) year, as there was no previous assessment used in clinic to be compared as a gold standard. At this point consequential evidence is largely theoretical, given that this pilot project was not designed to test the questions raised. However preliminary evidence that the tool caused raters to give increased formative feedback in clinic, and residents to change their behavior and ask for it, are great examples of consequential evidence.

Research ethics board approval. Research Ethics Board (REB) approval was sought at the Ottawa Hospital and granted on March 26, 2014 (Appendix G). An amendment to include feedback sessions after the pilot study was approved on September 3, 2014 (Appendix H).

Grant support. A grant application was successfully submitted to the University of Ottawa Department of Surgery Grant for the 2014-2015 year. This money contributed to research assistant support, printing and stationary, food for focus groups, and honoraria for participating residents. At the one-year mark a progress report was submitted to the Department of Surgery as an update on the project.

Steps of Tool Development

Step 1: Determining the purpose of the assessment tool. The first step in development of a tool should always be to determine if its purpose is formative or summative. The primary researcher, and a research team (on which there were four representatives from different surgical divisions, as well as three representatives of the educational administrative bodies) had early discussions directly addressing this topic. It was determined that, given

competency-based requirements in the near future, a formative tool was more necessary at this time.

Preliminary test specifications were also determined a priori (Downing, 2006, p.9). The type of testing format to be used was a combination of selected response using a rating scale and short answer comments (constructed response) to maximize feasibility. The total number of test items was flexible, however an assessment form greater than 2 pages would likely be infrequently used. Therefore the goal was to identify fewer than 20 items. The CanMEDS roles were considered for use as a cognitive classification system for the test blueprint, since the primary researcher was interested in observing how items overlapped with these competencies (see further explanation in *Step 3*). No visual stimuli (photographs, graphs, or charts) were planned for inclusion in the items and no scoring rubric was created due to the formative nature of the tool. There was no time limit set for completion, but the hope was that for feasibility reasons it would take less than 10 minutes to complete.

Step 2: Identifying construct of interest & content validity.

Literature review. An initial literature review was performed of published English language journals, with the assistance of two University of Ottawa Librarians and 1 librarian from the Ottawa Hospital. Using combinations of the following search and MESH terms, assessment methods of surgical clinic competence were investigated: “surgery”, “clinic”, “competence”, “competency”, “surgical”, “clinical”, “education”, “competence/standards” (MESH term), “assessment tool”, “tool”, “assessment”, “workplace-based assessment”, “surgical clinic”, “non-technical skills”, “soft skills”, and “intrinsic skills”.

Step 3: Review with content experts.

Introduction. Two consensus groups of expert surgeons, residents, and program directors with an interest in medical education were planned to gather content evidence. The goal was to develop a list of qualities and skills a resident should display in clinic before a staff surgeon will allow them to work independently as a staff surgeon. Alignment of these learning objectives and assessment is a key feature of any assessment tool and a basic starting point.

An option was to begin with the CanMEDS competencies, given that they are the RCPSC objectives for training. However, the goal of this project was to make the tool as applicable to surgeons as possible, and incorporate language they use on a regular basis in the clinic to describe their residents (see *Entrustability Decision Making*). Therefore, surgeon stakeholders

were approached first and asked what qualities they believe a competent surgical resident would display in clinic, allowing for the possibility that CanMEDS terminology would appear during this process.

The consensus approach used was closest to the Nominal Group Technique (NGT). As suggested by Humphrey-Murto et al (*manuscript in preparation*), the process is described here in detail to provide a comprehensive justification of the steps taken (Figure 2).

Figure 2: Checklist for Consensus Group Technique (Humphrey-Murto et al., <i>manuscript in preparation</i>)	
Were clear objective(s) described?	✓
Was a literature review carried out?	✓
How were the items selected for inclusion in the initial questionnaire?	✓
Was the process used justified? Was it described in appropriate detail?	✓
Was the most appropriate consensus method selected?	✓
How was the expert group defined and selected? Was a rationale provided?	✓
If NGT- was the choice of facilitator described?	✓
Was there a clear outline/description of the process involved?	✓
Was anonymity maintained?	No
Was feedback provided after each round? What type of feedback was provided?	✓
Was iteration used? Was the number of rounds decided a priori?	✓
Was response rate reported and did the discussion address potential implications? (<i>not necessarily applicable to NGT</i>)	N/A
How was consensus defined? Was it defined a priori?	✓
Did the discussion address potential methodological issues? Were the results credible?	✓

Steps used in the consensus group process. The nominal group technique (NGT), a sub-type of consensus group, is used to generate and prioritize ideas in consensus. The steps described here are specific to NGT. In this structured group interaction, a facilitator presents a key question producing answers that are able to be ranked (Murphy et al., 1998).

Prior to this consensus group, the problem was defined. The objective of the group was to create a list of key features a resident must display in surgical clinic to show independent competence. Purposive sampling was performed to gather a group of surgeons with medical education experience at an academic hospital (Jones & Hunter, 1995; Sargeant, 2012). This sampling involved all the divisions of surgery at the Ottawa Hospital. The goal was to recruit at least ten surgeons and two resident representatives (for two consensus groups of six people each). Ideally the participants would embody as many surgical divisions as possible, involve various levels of surgical education experience, and have had practical experience with assessment of surgical residents (Tammela, 2012). Three residents ended up participating in the NGTs, two from orthopedics and one from general surgery. These residents had the important role of presenting the perspective of those the assessment tool will be used to evaluate, and so their opinion was important.

Surgeon participants had from four to 18 years clinical academic hospital experience and included representatives from the following surgical divisions: general surgery, pediatric orthopedic surgery, orthopedic surgery, thoracic surgery, urology, obstetrics/gynecology, and vascular surgery.

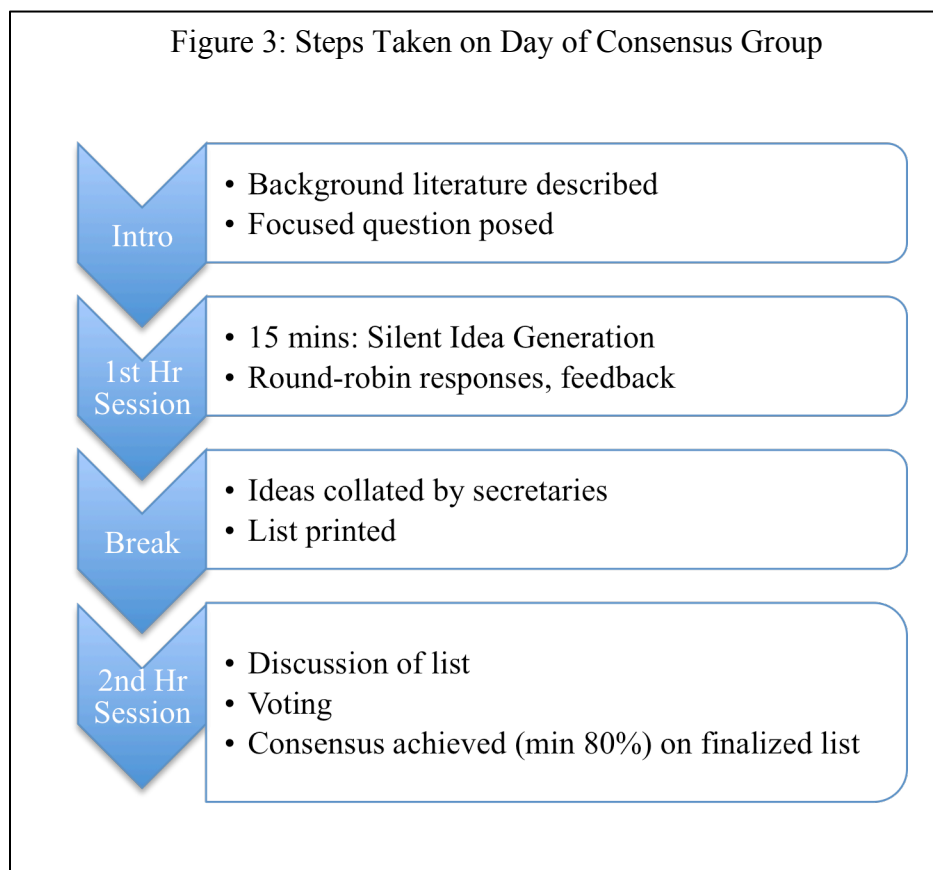
A facilitator (N.D.), who had previous experience leading focus and consensus groups, was asked to lead. This person was aware of, and involved in, developing the objectives and protocol of the research project. The primary researcher was present for the consensus groups to take notes and document, but did not lead for fear of a conflict of interest (personal relationship with some of the consensus group participants).

Consensus group one. On the day of the consensus group, participants were introduced to the purpose of the project and, to focus them, were asked a specific question. The question was, “What features would you like to see in a resident who is prepared to run a surgical clinic independently?” A few minutes were given for silent idea generation, before opening up the floor for group discussion. This is one of the hallmarks of NGT (Murphy et al., 1998). Once this time was up, the facilitator went around the semi-circle and asked each group member to respond with one answer. One secretary (primary researcher) recorded each idea in a succinct phrase on an easel visible for all participants to see while a second secretary (C.S.) took detailed written notes.

The floor was open to discussion of suggested answer items only after everyone had a chance to present their ideas (Campbell & Cantrill, 2001). To stimulate further discussion, the

participants were probed to envision qualities of a good resident (positive qualities) and qualities pointing to a poorly performing resident (negative qualities) (List, 2001). This process went on until the ideas started overlapping, or the participants ran out of ideas. The items were recorded by the research team and presented back to the group in aggregate format after a short break. During the break, the facilitator and secretaries typed and collated the ideas into a list and printed one list for each participant.

Following the break, participants were asked to comment with further ideas on each item on the list. The moderator asked the group to quietly rank the items thinking in terms of: (1) definitely must keep (2) not sure (3) get rid of. Then voting commenced, by raising hands, to decide if each item should remain on the list, with a consensus of 80% required for an item to remain. Of note, consensus was defined a priori as 80% agreement (considered strong agreement by the research team) on a particular outcome, or in practical terms, five of the six group members agreeing (Krueger & Casey, 2014). Then the facilitator asked the group to consider how the items lumped, and whether or not they would re-word any item that was recorded. See Figure 3 for a flowchart of the consensus group process.



Consensus group two. The second consensus group was run similarly. However, during the break, the list generated in the first half, and the list developed in the first consensus group, were combined into one list by the study team. The task of the second part of the consensus group then was to simplify the two lists into one (34 items) and identify similarities and differences. Ultimately, members voted on each item leading to one list of items. Items were eliminated from the list if 80% consensus was not achieved. The decision had been made a priori that if this final list contained more than 20 items (too long for an assessment tool), that the participants would be asked to rank-order their choices from most important to least important for inclusion. At the end this consensus group however, the list contained 13 items and this step was not necessary.

Again, voting was performed by holding up hands, and therefore complete anonymity *was not* as maintained. However participants were given the option to submit a ranked/revised list after the group (either in person or by email) if they felt their opinions had not been fully represented. No participants elected to do this.

National external expert survey. In order to bring in opinions from outside Ottawa, an online survey using intentional sampling of Surgical Educators from across Canada was performed. Surgical education experts from across Canada were identified by polling three surgical education experts in Ottawa for contacts (names and emails). This wide geographic sampling provides further evidence that the construct had been sampled.

An email was sent to the national experts explaining the purpose of contacting them, and that their answer would help the primary author with her thesis project (see Appendix I for the email script). The question was written in *Survey Monkey* and phrased, “Your resident is in clinic with you today and you have been asked to assess them. What are the top five things you would consider to assess regarding their competence to run this clinic?” Five free-form blank answer spaces were given.

Information received from these survey participants was reviewed and classified into categories by the author (mapping back to individual consensus group items, or not yet mentioned). Then a discussion occurred between the primary researcher and two team members (S.H., W.G) ensure agreement. Responses that fell outside the responses already captured in consensus groups were listed and reviewed for potential inclusion in the pilot study tool.

Step 4: Item & anchor development. A meeting with the primary researcher and two team members (W.G., N.D.) was held to transform the list of items collected from the consensus groups into an assessment tool. They also discussed how to transform the list of features into items to be assessed on a six point rating scale (see Figure 4).

Originally, the anchor scale was five points. However one of the study team members is a fellowship Program Director (W.G.) and he felt strongly that if there was a six option on the scale dedicated to “fellowship level performance”, this tool could potentially be used for assessing fellows as well. All three group members were actively involved in this discussion process and a six-point scale was decided on.

A global assessment item and an entrustability-aligned anchor scale were modified from previous work at the study center (Gofton et al., 2012), to fit the construct and context of this tool. Scale descriptors were written based on discussion at the consensus groups as well as previous experience with entrustability scales. See Appendix J for the comparison of the original construct-aligned anchor scale and the OCAT anchor scale.

Figure 4: OCAT Entrustability Anchor Scale

Scale

- 1—“I had to do”—i.e., Requires complete guidance, unprepared to do, or had to do for them
- 2—“I had to talk them through”—i.e., Able to perform but requires repeated direction
- 3—“I had to direct them from time to time”—i.e., Demonstrates some independence, but requires intermittent prompting
- 4—“I needed to be available just in case”—i.e., Independence but needs assistance with nuances of certain patients and/or situations, *unable to manage all patients*, still requires supervision for safe practice
- 5—“I did not need to be there”—i.e., Complete independence, can safely and manage a general clinic in your specialty
- 6—“Functioning at level higher than a generalist in my specialty”—i.e., Fellow level or subspecialty practitioner

Pre-pilot project. The preliminary assessment tool resulting from this meeting was distributed to the consensus group participants (residents and surgeons). They were asked to review the tool and make suggestions. The surgeons were asked to try using it in their own clinic on a few trainees. They provided feedback regarding clarity and utility of the scale over two weeks by email. The primary researcher collated these email comments into one document with action items highlighted. The tool was revised based on these action items and feedback. The resulting OCAT was used in the pilot study.

Step 5: Pilot project

Identify participants. Pilot study participants included resident trainees and staff surgeons in the divisions of orthopedics, obstetrics and gynecology, and general surgery at the Ottawa Hospital. This included three campuses in Ottawa: the Civic, General, and Riverside campuses. Clinics could be either generalist or specialist clinics, and the only requirement qualifying a clinic for the study was that a resident was working in it with staff supervisor. The expected sample size proposed for the three-month pilot was 25 residents rotating through the orthopedic, obstetrics and gynecology, and general surgery clinics (accounting for some residents being on rotations without clinic exposure).

Step 6: Rater orientation sessions. Staff raters and resident orientation sessions were held in each of the three divisions, excluding the raters in the obstetrics and gynecology division. These sessions were incorporated into the beginning of academic half day (residents), grand rounds (staff and residents), and divisional meetings (staff). Unfortunately, it was not possible to coordinate a meeting with the staff raters from obstetrics and gynecology. An email introduction outlining the concepts reviewed in the orientation-session was sent out to these staff instead. Not all staff and resident members from all divisions were present at the training sessions despite efforts by the primary researcher to contact them.

Staff and resident orientation sessions included a ten-minute didactic presentation introducing the following information: the vision of the study (why they should participate), logistics of the study (e.g. where to turn in OCAT forms), and what their consent to participate meant. The assessment tool items and the entrustability scale were demonstrated. Staff surgeons were reminded of the importance of timely, constructive feedback, and the importance of identifying the trainee requiring remediation. They were also reminded that the tool is an adjunctive assessment score, and similar to the mini-CEX (Norcini et al., 2003) and O-SCORE (Gofton et al., 2012) not meant to be used alone. Similarly, trainees were educated that they should not expect to be evaluated as competent early in their training.

Prior to these presentations, an email was sent out to briefly introduce the study and to let staff and residents know the orientation session would be held. See Appendix K and L for copies of the staff and resident recruitment emails (sent by the study research assistant C.S.). Following the OCAT orientation session, a follow-up email for reference to topics covered was provided.

Consent process. Potential participants present at the orientation sessions were asked to consent to the study if they agreed with its terms. See Appendix M for a copy of the consent form. The consent forms were either collected from them that day, or they were given the chance to email them to the study research assistant. It was clear to the participants that this study would not reflect on his or her academic standing (residents) or their employment (staff raters) and that all results would be kept in a confidential and de-identified manner. A master list of participant names, matched to study ID was kept on a password-protected computer in the research assistant's (C.S.) office.

Pre-survey question. Any resident who consented to participate was given the option of completing a one question pre-survey to be turned in with their consent form. Their responses were de-identified. The pre-survey question was, "I receive appropriate feedback regarding my performance in clinic at the end of most clinics I participate in". Residents were asked to give their answer on a five-point scale of 1- strongly disagree, to 5- strongly agree. There was a blank space provided to prompt "*Comments*". See Appendix N for the pre-survey.

Six-month pilot study. The assessment tool was made available in paper form in the main dictating/reporting area of each clinic (with permission from attending surgeons and clinic staff). When a trainee participated in a day or a half-day of clinic, they had the option to ask their supervising surgeon to fill out a form. Nothing was mandatory. The pilot study originally was planned for three months (July-September), but given a summer slow-down in clinic attendance, the study was extended for three further months (October-December) to collect an adequate number of OCAT forms.

Throughout the six-month period, the primary study investigator spoke to residents and staff surgeons to continually gather informal feedback. This information was recorded by date in a word document and any repeating comments were noted to discuss with the full study team on an ongoing basis. One piece of informal feedback that recurred a few times was an interest in the six-point scale, and the difference between the PGY5 level and the fellow level. Based on this feedback, halfway through the study (at the end of October), a vertical line was added to the tool between anchor levels five and six. See Appendix O the OCAT form containing a vertical line. This change was noted informally to make a positive difference so no further action was taken.

Psychometric data analysis. Data were recorded in an excel file on an on-going basis throughout the study. The primary investigator went through entire excel file a second time following closure of the pilot study to ensure data quality, and check for incorrect entries. It was decided that items 12 and 13 (technical skills questions) were unique enough (and visually separated by a horizontal line on the form) that they should be treated as a subscale in the analysis. The global assessment item was similarly analyzed separately. Following data entry, the data was cleaned. The following data cleaning steps were taken:

1. Started with a total of 155 OCAT forms collected.
2. Despite rater training, infrequently raters tried to utilize a 0.5 score. All 0.5 scores were rounded up to the nearest whole number value (ex. 4.5 to 5). This only occurred on 10 items.
3. Removed one form that was a clear outlier (a PGY-1 filled out his/her own form and gave himself all 6's with no staff signature on the form).
4. All entries with missing data points in items 1-11 were removed (total forms removed 12). See Table 1 for missing data frequency by item.
5. Removed items with no supervisor or trainee study ID (supervisor 0, resident 1).
6. Removed data from the 1 fellow that participated (total forms removed 9).
7. Remaining OCAT forms = 132.

Table 1: Missing Data Frequency

Missing Data Frequency:	
Q1	2
Q2	1
Q3	2
Q4	2
Q5	3
Q6	0
Q7	0
Q8	0
Q9	5
Q10	1
Q11	0
Q12	93
Q13	84

Scale development can follow one of two traditions within psychometrics; classical test theory or modern test theory with a focus on item response theory (IRT). IRT has certain advantages over classical test theory, such as: not assuming ordinal data can be treated as interval scales, more precise measurement of error (based on the individual, not the group, as in classical), facilitation of comparison across tests, and ability to have varying numbers of options on a rating scale within the same assessment tool. However, classical test theory does not require the large sample size typically associated with item response theory, can be applied in almost all measurement situations due to its soft assumptions, and is more commonly used in medicine and therefore produces statistics that are more collectively understood and explained (De Champlain, 2010). Therefore, while item response theory definitely offers some advantages over classical test theory, there is nothing definitively wrong with classical test theory as it applied well to the setting of the OCAT assessment tool for the reasons listed.

Descriptive statistics were calculated for each item as well as internal consistency. In addition, item-total correlations and inter-item correlations were calculated to look for highly related or unrelated items. A generalizability analysis (g study) was used to determine the

reliability of the scale with participants (residents) nested within surgical specialty and crossed with items. The advantage of a generalizability analysis is that it uses analysis of variance (ANOVA) to portion out variances associated with all the factors used in the model. Although there was some evidence of skewness in the item ratings, ANOVA models have been shown to be robust and can be used with non-normal distributions (Norman, 2010). The results of this analysis generated a reliability coefficient and allowed determination of differences between surgical specialties. In addition, as a further check on the validity of the scores, the scale totals were analyzed using training level as a factor to determine if scores increase as a function of training. Scale totals were also analyzed against the function of the clinic in which the assessment was performed; i.e. generalist clinic or specialist clinic. Partial eta square effect sizes were calculated throughout (Sullivan & Feinn, 2012) using the following interpretation scale: from 0.02 to 0.13 is a small effect size, 0.13 to 0.26 is medium, and greater than 0.26 is a large effect size.

Analysis of short answer questions. Summative content analysis (Hsieh & Shannon, 2005) was applied to the answers from the two short answer questions at the end of the OCAT. The primary researcher and the research assistant (C.S.) developed a coding taxonomy, taking into account which answers borrowed direct wording from items 1-11 (e.g. Item 7: Patient & Family Communication). See Appendix P for the coding strategy for question 14 & 15. All answers were coded independently; following which consensus was achieved on all answers. Absolute values were also calculated to determine how many of these questions were completed by staff surgeons.

Frequency of direct correlation to each item was calculated, as well as the tone of the answers (positive, negative). If answers did not lump into one of the first eleven items, the comments were subsequently classified based on specificity as global (e.g. “great job”) or descriptive (e.g. “good physical of a complicated APR patient”) (Espinell, Shah, Beach, & Boss, 2014). When an attribute outside of the initial coding template was mentioned, it was either integrated or noted as an outlier for later comment.

Step 6B: Feasibility assessment - Feedback sessions & survey. In order to guide further study and modify the tool for subsequent iterations, formalized feedback sessions were held. The assumption was that higher quality feedback would be received in this forum compared with an email survey. Two purposefully sampled groups (one of residents and one of

staff surgeons), who had been registered as participants of the pilot study, were recruited. Prior to the feedback sessions, the primary researcher wrote questions with input from the study team (see Appendix Q). During the sessions, questions were posed to the group by the same facilitator that ran the consensus groups (N.D.), and clarifying questions were posed if an answer was not clear. Each session lasted approximately an hour each. The primary researcher and research assistant (C.S.) took notes during the session, and the sessions were recorded for reference to specific quotes after. Five surgeons participated in the staff session and six residents in the resident session. Every participant was asked for his or her opinion on each question, and if someone voiced an opposing viewpoint it was heard. Answers were collated after the sessions by the primary researcher, the facilitator (N.D.), and the research assistant (C.S.) during a two-hour meeting, and results of this discussion were used to interpret the psychometric data.

Feedback survey. All participants were given the opportunity to provide written feedback on the clarity and utility of the scale through an email survey. The survey contained the same questions asked in the feedback sessions. This survey was not mandatory, but gave each participant the opportunity to express his or her opinion. In a few cases, staff surgeons and residents were interviewed about their survey answers in a one-on-one format with the primary investigator. In this way, opinions on the pilot project were solicited from as many study participants as possible.

Section 4: Results

Introduction

Reporting of the results will follow a similar pattern as the methodology, using Hamstra's seven steps of test development as an outline (2012). In this section, individual OCAT items will be referred to by their corresponding question number (Q1, Q2, etc). Please see Table 2 for an easy-access glossary.

Table 2: Glossary of Items on the OCAT

Item Number	Item Wording
Q1	Clinical Preparedness: <i>Arrived at clinic with appropriate knowledge base for anticipated patients</i>
Q2	History: <i>Efficient data gathering</i>
Q3	Physical Exam: <i>Efficient and accurate examination</i>
Q4	Case Presentation: <i>Synthesis of history and physical, clear presentation</i>
Q5	Differential Diagnosis: <i>Able to make a diagnosis and appropriately consider alternatives</i>
Q6	Management Plan: <i>Able to develop relevant plan dependent on context and be decisive (i.e. appropriate investigations, procedures, etc.)</i>
Q7	Patient/Family Communication: <i>Effective, sensitive, and respectful communication skills (verbal & non-verbal), language appropriate to patient understanding, able to build rapport and trust</i>
Q8	Documentation within Clinic: <i>Orders, prescriptions, forms, etc. (may not include consultation report)</i>
Q9	Collaboration: <i>Works well with and/or teaches other team members as appropriate (ie staff, student, other healthcare professionals)</i>
Q10	Time Management: <i>Able to economize time, manage interruptions, and modify time spent with individual patients appropriately</i>
Q11	Attitude: <i>Engagement and ownership in running clinic</i>

Sub Scale	If technical skills were performed in clinic
Q12	Technical: <i>Safely and effectively performs appropriate clinical procedures</i>
Q13	Situational Awareness: <i>Non-technical aspects of procedure (i.e. insight into patient experience, respects patient comfort)</i>
GAI	Global Assessment Item <i>Resident is safe to independently manage a general clinic in your specialty</i> Y/N
Short Answer	
Q14	Give at least 1 specific aspect of clinic done well
Q15	Give at least 1 specific suggestion for improvement

Steps of Tool Development

Step 1: Determine the purpose of the tool. It was first determined that the OCAT is to be used for formative purposes. There was no plan to set a pass score, as would be required by a summative tool. Also, the goal of the tool is to enhance feedback to residents on a daily basis making a formative purpose ideal. See results of the feedback session below for staff and resident response to the effect of the OCAT on formative feedback. In addition to enhancing feedback, additional goals of the tool were for it to:

- Demonstrate multiple sources of validity evidence, including reliability.
- Be feasible to use in a surgical workplace for assessment.
- Show high usability for surgeons assessors.

Step 2: Content validity evidence.

Literature review (construct of interest). The construct of interest for this tool is: surgery resident independent clinic competence over a day of clinic. Assessment tools addressing this particular construct were not available in the literature.

Identify stakeholders. A comprehensive list of stakeholder opinions to consider was written, including:

- Residents
- Staff Surgeons
- Patients

- Accreditation bodies (RCPSC, etc.)
- Hospital administration
- Program Directors
- Department of Surgery

Step 3: Review with content experts. Stakeholders were identified at the local and national level. Locally surgical educators and residents from all divisions of surgery were targeted. Surgical divisions represented in the consensus groups included: general surgery (4 persons), orthopedic surgery (2), obstetrics & gynecology (1), urology (2), thoracic surgery (1), pediatric orthopedic surgery (1), breast surgery (1), and hepatobiliary surgery (1). Resident representation was from general surgery (1) and orthopedic surgery (2).

National level surgical experts were identified to complete a survey from the following ten academic institutions: McGill University (Montreal), University Health Network (Toronto), Hospital for Sick Children (Toronto), St. Michael's Hospital (Toronto), McMaster University (Hamilton), University of Saskatchewan (Saskatoon), University of British Columbia (Vancouver), University of Calgary (Calgary), and unknown (anonymous). They were from a variety of surgical specialties.

Content validity evidence collected.

Results of consensus groups.

Summary of consensus group one. There was an initial focus in this consensus group on non-technical skills (communication, reliability, ability to manage time, etc.). This group determined that other tools already in circulation could assess technical skills. There was a discussion of how professionalism can mean various things to different people and is very multi-faceted. Appearance (dress, carry, etc.) is only one component. The word empathy was discussed in the context of recognizing the patient's emotions, as well as employing empathy while breaking bad news. It was agreed that the specifics of documentation were not necessary, but that quality record keeping is an important skill for a competent surgeon to learn (likely best learned in clinic). The group also had a discussion of qualities defining a poor resident in clinic versus a great resident (see Table 3).

Table 3: Consensus Group Resident Qualities Discussion

‘Poor resident qualities’	‘Great resident qualities’
Slow	Initiative
Creates controversy, upsets family and/or patient	Sharing knowledge from other clinics with staff/other trainees
Incorrect plan	Did the resident teach you?
Poor communication with nurses	Teaches medical students appropriately (within time limits)
Inability to convey concise/accurate history	Overall is a professional
Inappropriately dressed	Maintains confidentiality
Arrived late	Maintains composure

At the end of the first consensus group, there was a list of 12 items (Table 4) defining the construct of interest, not ranked in any particular order.

Table 4: List of Items Following First Consensus Group

1. Overall time management of clinic
2. Effective communication skills with patients
3. Quality of documentation
4. Ability to present case
5. Ability to perform focused history
6. Ability to perform a focused physical exam
7. Ability to develop appropriate management plan
8. Clinic preparedness
9. Manages interruptions
10. Ability to teach as appropriate
11. Collaboration with interdisciplinary staff
12. Initiative in running clinic

Summary of consensus group two. The second local consensus group generated an independent list of items, and then compared their list with the list generated in the first group. Specific longer discussions surrounded documentation and technical skills assessment, as they disagreed with the first group on these. On the topic of documentation, it was determined that dictated/transcribed notes should not be assessed at the same time as the clinic since not all residents dictate during the clinic (and they were not sure this should be encouraged either).

As for technical skills, they argued that there are some clinics in which many procedures are done (in other specialties that were not represented in the first consensus group). The non-technical skills involved in procedures are rarely assessed (e.g. situational awareness of performing a procedure on an awake patient), and that should be included in this tool. Following the second consensus group, there was a combined list of 12 items (Table 5).

Table 5: List of Items After Second Consensus Group

- 1- **Overall time management of clinic:** economy of time, showing up on time, manages interruptions, prioritization of patients, manages interruptions
- 2- **Effective & sensitive communication skills with patients:** effectiveness of the communication, using proper terminology and language with the patient, demonstrates respect for patient, sensitive communication (bad news, informed consent, during procedures), teaching patients effectively, ability to build rapport leading to trust, demonstrates ability to respond appropriately to verbal and non-verbal cues
- 3- **Clinic preparedness:** knowledge base, pre-knowledge of patients
- 4- **Collaboration with interdisciplinary staff:** professionalism, teach as appropriate (i.e. staff, student, other healthcare professional)
- 5- **Engagement of resident in running the clinic:** ownership, attitude, confidence
- 6- **Quality of documentation:** clarity and appropriateness (consider having an N/A option, group was concerned about feasibility issues)
- 7- **Ability to perform focused history:** data gathering, initial evaluation
- 8- **Ability to perform a focused physical exam**
- 9- **Ability to synthesize information and present a case:** synthesis of story, art of retrieval, ability to apply their knowledge to the patients seen in clinic
- 10- **Ability to make a diagnosis/differential diagnosis:** pattern recognition
- 11- **Ability to develop appropriate & relevant (necessary) management plan dependent on context:** ownership, decision making, how to order tests in such a way that the most effective and useful information is yielded/ how to order (manager role), productive/relevant investigations
- 12- **Situational awareness** when performing clinic-based procedures:
 - a. 2nd group thought there is a distinct difference between procedures performed under general anesthesia and those performed under local
 - b. Nontechnical procedural skills - knowing your environment, insight, respect, situational awareness when performing technical skills, patient discussion, consent, patient comfort, lighting, etc.

External surgical education expert survey. The national survey had a response rate of 83%. A purposive sample of 18 surgical education experts from a variety of surgical divisions (see Table 6) was emailed and 15 responded (87%). A potential source of bias resulted as the sample included three respondents from the same hospital (University Health Network in Toronto), and three who wished to remain anonymous, meaning they could also be from Toronto. However, the sample did include wide geographic representations from as far east as Montreal, and as far west as Vancouver. Additionally, Toronto is a national center for surgical education with a large population of trainees, and so it could be argued that a larger representative sample from that location is appropriate.

Table 6: Academic Affiliations of National Expert Survey

Academic Affiliation	City	Number of Respondents
McGill University	Montreal, QC	1
University Health Network	Toronto, ON	3
Sick Kids Hospital	Toronto, ON	1
St. Michael's Hospital	Toronto, ON	1
McMaster University	Hamilton, ON	1
University of Saskatchewan	Saskatoon, SASK	1
University of British Columbia	Vancouver, BC	1
University of Calgary	Calgary, AB	1
Anonymous	Unknown	3
Total		15

The answers generated in the local consensus groups were compared to the survey responses from the national experts to see how many agreed (Table 7 and 8). Interestingly, the majority of survey responses correlated directly with items already summated from the focus groups. The remainder were either comments that fit into previously established items (e.g. 'initiative, not requiring too much hand-holding' fit into attitude/ownership of clinic) or interpreted as too broad to be directly incorporated (e.g. 'good judgment'). See Table 7 for more information.

Table 7: National Expert Survey Responses
(bold = matches consensus group item)

# of Answers Correlating to Consensus Group Items	5 Answers Given
4/5	1. Ability to take a relevant history. Not a comprehensive history, a history that relates to the problem at hand. 2. A detailed, accurate examination of the affected anatomy. 3. Ability to formulate a reasonable treatment plan given the information gathered from history, physical and imaging. 4. Don't be lazy! Do what it takes to get the relevant information and do what it takes to provide the best treatment. (Attitude) 5. Do not delegate important tasks to others, regardless of your level of training.
4/5	1. Ability to dictate the clinic note number one (documentation) 2. Professional attitude and appearance (Attitude) 3. Keen and interested (Attitude) 4. Ability to synthesize info in a timely (and therefore helpful) manner (Case presentation) 5. Initiative, not requiring too much hand holding. - attitude
6/7	1. Punctuality - must be on time or early to avoid clinic delays and problems with patient flow 2. Prepared for clinic - read around complex cases and to review relevant anatomy - very relevant for specialty clinics that demand expert knowledge and skills in assessment (Preparation) 3. History taking skills 4. Physical examination skills 5. Verbal reporting on findings to attending surgeon - must speak to relevant history and PE findings along with treatment plan 6. Professionalism - introduction, how they handle themselves with the patient and their family, effective communication, etc... 7. Ability to work with team of health care professionals
3/5	1. Decision making skills 2. Time management skills 3. Whether resident is post call 4. Workplace attitude 5. Triage (included in time management)
5/7	I have residents and fellows essentially every week in Clinic and I doubt I have anything original to add for you, but in no particular order: 1. Punctuality, and "professionalism" with Clinic staff and with patients; 2. Ability to take a good history and physical; 3. Ability to read the imaging properly; 4. Ability to formulate a diagnosis and treatment plan that are reasonably accurate; 5. Time management.
	"Good question but often needs context to answer completely. I am looking for different things in a PGY 1 vs PGY 3 or 5. Often stress basic skills with completeness for junior and look for more organization and management with more senior trainees. Might be better to frame your question with a PGY level of resident (or basic rotation vs advanced for competency based programs)."
2/5	1. Knowledge (Preparation) 2. Clinical judgment 3. Clinical experience 4. Understanding of how to individualize patient care based on understanding of the individual patient (Communication – to patient level) 5. Understanding of the system of medical care delivery
5/7	1. Ability to take a concise history, physical examination, w/up and plan 2. Ability to review patients efficiently and effectively 3. Ability to communicate with patients 4. Ability to demonstrate professionalism (dress code, attitude and behaviors) 5. Ability to manage diverse patient populations
3/5	1. Knowledge of disease processes / natural history of presentations in my area (preparation) 2. Appropriate interpersonal communication skills 3. Insight into own abilities 4. Data synthesis skills - putting Hx, PE, and lab/imaging results together coherently 5. 'Trustworthiness'

# of Answers Correlating to Consensus Group Items	5 Answers Given (Continued)
5/6	1. Ability to take a "focused" history 2. Ability to accurately perform "focused" physical 3. Synthesizes clinical problem and communicates this effectively 4. Development of reasonable treatment plan 5. Professional behavior including attire and appearance 6. Time efficiency
3/5	1. History taking (organized, thorough, complete) 2. Physical examination (organized, thorough, complete) 3. Interpretation of x-rays and other tests 4. Ability to communicate with patient 5. Professional role Additional comment: varies with level and experience of resident
5/5	1. Appropriate treatment plan 2. Ability to make an accurate diagnosis 3. Interpersonal skill 4. Thorough history taking 5. Thorough physical exam
4/7	1. Ability to create a differential diagnosis and 2. Provide supporting history and physical 3. Efficiency with patients. 4. Being able to perform a focused/relevant history and 5. Physical 6. Appropriate charting 7. Interpretation of imaging
5/5	1. Positive attitude and pleasant demeanor 2. Time management: taking more time with patient when it is needed 3. Focus on chief complaint but accrue all necessary information (history) 4. Management of human resources: nurse(s), cast tech, clerks, residents, etc (time management) 5. Adequate knowledge and experience (preparation)
5/6	1. History taking - correct info, streamline gathering 2. Physical exam - general and focused 3. Professional attitude - appearance 4. Communication skills with my staff and patients 5. Good differential and 6. plan to identify and address diagnosis

When the survey answers were mapped to the items developed in the consensus groups (Table 8), all of the first 11 items were mentioned at least once on the survey. The two technical skills items were not mentioned, but were kept in due to the insightful discussion in the second consensus group.

Table 8: Analysis of Responses from National Survey Compared to Consensus Groups

Consensus Group Items	Survey Responses that Agreed	Interesting Comments/Ways that people said it
Clinical Preparedness (knowledge base for anticipated patients)	2	“Prepared for clinic - read around complex cases and to review relevant anatomy - very relevant for specialty clinics that demand expert knowledge and skills in assessment” “Knowledge” – top hit in one response “knowledge of disease processes / natural history of presentations in my area”
History (efficient data gathering)	6	“Not a comprehensive history a history that relates to the problem at hand.” “focused”
PE (Efficient & accurate)	6	“A detailed, accurate examination of the affected anatomy.” “focused”
Case Presentation (Synthesis, clear)	4	“ability to synthesize info in a timely (and therefore helpful) manner” “ability to review patients efficiently and effectively” “data synthesis skills - putting Hx, PE, and lab/imaging results together coherently”
Differential (Diagnosis and consider alternatives)	1	
Management Plan (relevant plan including imaging, procedures, etc) considering context, decisiveness	5	“Ability to formulate a reasonable treatment plan given the information gathered from history, physical and imaging.”
Patient/Family Communication (sensitive, effective, respectful- verbal & non-verbal, language appropriate to patient level, rapport and trust)	5	“Professionalism - introduction, how they handle themselves with the patient and their family, effective communication , etc.” “ understanding of how to individualize patient care based on understanding of the individual patient”
Documentation (may not include consultation report)	1	“ability to dictate the clinic note number one”
Collaboration (team members)	1	
Time Management (economize time, manage interruptions, modify time with patients)	3	
Attitude (engagement & ownership in running clinic)	3	“Don't be lazy! Do what it takes to get the relevant information and do what it takes to provide the best treatment.”
Technical Skills (safely & effectively performs)	0	
Technical (Situational Awareness, insight into patient experience, respects patient comfort)	0	

Recurrent survey answers that were not mentioned in the consensus group were reviewed for possible inclusion into the OCAT tool (Table 9). The most stark finding was that six experts surveys mentioned ‘professional attitude and/or dress’ in some manner. Potentially this falls

under the ‘attitude’ item, however the primary researcher thought it should be revisited after the pilot project. Professionalism has many definitions, and the meaning of the word can be open to interpretation on a very individual basis. Therefore for clarity, the word ‘professional’ was left off the OCAT pilot version, but noted to be asked about in the feedback sessions.

One further idea brought up was the word ‘trustworthiness’. The entrustability rating scale is built on this concept, and so the idea of trust is built into the structure of the OCAT. It was reassuring to see that this surgical education expert considered it valuable enough to be in their top five characteristics of resident performance in clinic.

Table 9: Responses from National Survey Not Appearing in Consensus Group List

Answer on Survey	Decision Made by Team
Do not delegate important tasks to others, regardless of your the level of training	Difficult to word this. Falls under teamwork. Do not include.
Ability to dictate the clinic note as number one.	Decided this falls under documentation item.
Professional attitude and appearance (x 6)	Only mention that is close to this in items is attitude. Debated whether professionalism should be an item on the tool. Determined that definition of professionalism is too variable. Left off tool for now. Will re-visit after pilot project.
Keen and interested	Falls under attitude
Initiative, not requiring too much hand-holding	Falls under time management & attitude
Punctuality	Decided not to include this word, but did believe it fell under our clinic preparedness definition.
Decision making skills	Included under management plan item (decisiveness)
“Whether resident is post call”	Did not know what was meant by this. Comes into environmental biases affecting the tool.
Triage	Falls under time management. Will re-assess after pilot project if triage should be stated explicitly under time management category.
Ability to read imaging correctly x 3	Did not create a category for this. Falls under the management plan (synthesizing). Will reassess after pilot project.
Clinical Judgment & Clinical Experience	Did not include this because team believed it was too vague. Falls into a few pre-existing items.
Understanding of the system of medical care delivery	Interesting. Think this falls into time management.
Ability to manage diverse patient populations	Interesting also. Wasn’t included as an item because team couldn’t understand what was meant by “diverse patient populations”
Trustworthiness	Like this idea. Started a discussion about entrustability and how this concept will be included in the OCAT already through the entrustability anchor scale. Do not need an item specifically for it.

Step 4: Item development. Descriptors were written during a two-hour meeting between the primary researcher, and two study team members (N.D. and W.G.) based on discussion in the consensus groups, informed by survey answers. A list of the final items included in the OCAT tool can be found in Table 10.

Table 10: Final List of Items and Descriptors

Item	Descriptor
Clinical Preparedness	<i>Arrived at clinic with appropriate knowledge base for anticipated patients</i>
History	<i>Efficient data gathering</i>
Physical Exam	<i>Efficient and accurate examination</i>
Case Presentation	<i>Synthesis of history and physical, clear presentation</i>
Differential Diagnosis	<i>Able to make a diagnosis and appropriately consider alternatives</i>
Management Plan	<i>Able to develop relevant plan dependent on context and be decisive (i.e. appropriate investigations, procedures, etc.)</i>
Patient/Family Communication	<i>Effective, Sensitive, and respectful communication skills (verbal & non-verbal), language appropriate to patient understanding, able to build rapport and trust</i>
Documentation within Clinic	<i>Orders, prescriptions, forms, etc. (may not include consultation report)</i>
Collaboration	<i>Works well with and/or teaches other team members as appropriate (i.e. staff, student, other healthcare professionals)</i>
Time Management	<i>Able to economize time, manage interruptions, and modify time spent with individual patients appropriately</i>
Attitude	<i>Engagement and ownership in running clinic</i>
If technical skills were performed in clinic	
Technical	<i>Safely and effectively performs appropriate clinical procedures</i>
Situational Awareness:	<i>Non-technical aspects of procedure (i.e. insight into patient experience, respects patient comfort)</i>

Global assessment item. A global assessment item, modeled after the O-SCORE global assessment item (Gofton et al., 2012), was written. This item was phrased as: “Resident is safe to independently manage a general clinic in your specialty”.

Pre-pilot project. The OCAT underwent a two-week mini-pilot before the full pilot project began, utilizing the clinics of the surgeons that had participated in the consensus groups. Specific comments were received from five of the participants (four staff surgeons and one resident). One change was made to the tool which involved changing the wording of “general” clinic to “generalist in your specialty”. Another comment referred again to including professionalism (specifically appropriate attire and being on time). Given that this had already been addressed in the external expert survey and the consensus group, the decision to not include the word ‘professionalism’ persisted.

Finally, there was a request for consideration of a “Not Applicable” column alongside the rating scale in case items did not apply. Instead of adding a ‘not applicable’ category, the technical questions (Q12 & 13), which were most likely to be ‘not applicable’, were separated out from the remainder of the items by a statement (if procedures were performed in clinic) and horizontal lines surrounding them. The remainder of the pre-pilot comments approved of the tool, and none of the items on the tool were changed further.

Step 5: Pilot test the instrument.

Pre-survey of residents. Prior to tool distribution in clinics, resident participants were asked to complete a one-question survey along with their consent form. Their responses were made anonymous. The pre-survey question was, “I receive appropriate feedback regarding my performance in clinic at the end of each clinic I participate in,” with a five-option scale response scale from 1-Strongly Disagree to 5- Strongly Agree. There was also a box available for written comments which quite a few residents took advantage of.

A total of 80 residents consented to participate in the OCAT study (Table 11) and 52 of them completed the survey (65% response rate). See the response rate broken down by specialty in Table 12.

Table 11: Number of Residents Consented for OCAT Study by Division

Total # Residents Consented	80
Orthopedics	22
General Surgery	37
Obstetrics & Gynecology	21

The mean answer to the survey question implied residents do not feel they receive appropriate feedback. Average response value was 2.6 (between disagree and neutral to the question). A breakdown by discipline showed less feedback is being given in orthopedics and general surgery (averages of 2.4 and 2.3 respectively) compared to a 3.2 (skewed towards a positive response to the question) for obstetrics and gynecology.

Table 12: Pre-Pilot Survey Results - Resident Satisfaction with Current Feedback Rates

"I receive appropriate feedback after clinic"	Overall # of Responses	Ortho	Gen Surg	Obs&Gyn
1 – Strongly Disagree	7	4	2	1
2 - Disagree	19	3	13	3
3 – Neutral	14	2	6	6
4 – Agree	11	3	2	6
5 – Strongly Agree	1	0	0	1
Total Responses	52	23	23	17
Mean Overall Response to Question	2.6 (disagree to neutral)	2.4 (disagree to neutral)	2.3 (disagree to neutral)	3.2 (neutral to agree)
Percentage Response Rate to Survey	65%	55%	62%	81%

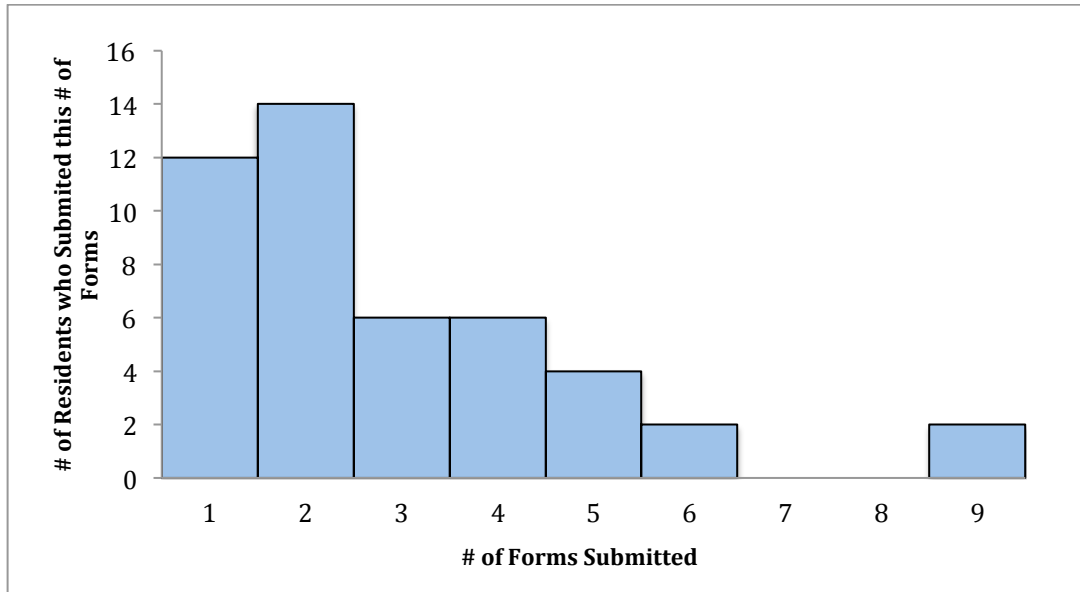
All the free-text comments received on the survey are listed in Table 13. They all seem to be given by residents who feel strongly about their lack of feedback, ranging from "I'm not sure I've ever had feedback in clinic" to "Highly variable depending on staff and clinic".

Table 13: Free-Text Comments Received on Resident Pre-Survey

Staff usually just say "thanks", rarely give specific feedback.
Usually receive a thank you at most. Instruction on what to improve would be nice!
Rarely, if ever, receive feedback in clinic
Minimal personal feedback, good feedback on how to assess patient presenting concerns.
I find it very difficult to receive timely feedback that is specific to how I did in clinic. Often feedback is generic and not useful to actually work on my clinical skills.
I'm not sure that I've ever had feedback after a clinic
Minimal feedback
I have done few clinic rotations, but have received appropriate feedback so far.
Highly variable depending on staff and clinic (i.e. more feedback after gynecology clinics vs. obstetrics clinics)
Most staff do not give feedback
I have done few clinic rotations, but have received appropriate feedback so far

Pilot study descriptive statistics. A total of 44 surgeons from one of three disciplines (orthopedics, general surgery, obstetrics/gynecology) used the OCAT to evaluate 46 residents over six months of clinic performance. Three surgical specialties participated in the OCAT study. A total of 24 observations were excluded due to incomplete information leaving 132 complete forms for an average of three (2.86) per resident. Figure 5 shows the distribution of how many forms were submitted by each of the 46 residents (minimum one form, maximum nine), therefore most residents were assessed more than one time in more than one clinic context.

Figure 5: Number of OCAT Forms Submitted by Individual Residents



The breakdown of which of the three participating specialties they were in can be seen in Table 14, including number of forms submitted per specialty.

Table 14: OCATs Submitted by Discipline

Discipline	Number of Complete Forms
General Surgery	65
Obstetrics & Gynecology	21
Orthopedic Surgery	46
Total	132

Psychometric analysis.

Item analysis. Table 15 shows the descriptive statistics for each item on the scale. As shown in the table, the mean rating for each item was relatively high (all means above 4.0 out of 6). However, for each item despite the relatively high means, there was a range of scores used with low scores of 1 or 2 assigned for each item indicating that the raters were willing to use the lower end of the scale. To determine if the distributions were relatively normally distributed (to indicate whether parametric or non-parametric tests should be used), the skew and kurtosis for each item was determined by calculating z-scores for each item, as shown in Table 15. If the scores contain significant skew or kurtosis, non-parametric tests should be used. Skew refers to

the amount of asymmetry of the distribution of scores around the mean. A normally distributed set of scores has a skew value of zero. Similar to skew, kurtosis is a measure of the shape of a probability distribution and refers to the amount of “peakedness” of the data, or the width of the peak. In Appendix R histograms depicting the scoring distribution of all the items (1-11) can be found. Items one through eleven all had distributions that were skewed, and four items (3, 6, 9, & 11) indicated that there was kurtosis in the item distributions. Most items showed negative skew (tail to left was longer) and kurtosis centered on a score of four or five

Table 15: Item-Level Statistics for the Ottawa Clinic Assessment Tool (OCAT)*

OCAT Item	N	Rating ^φ		Skewness		Z-score ^ξ	Kurtosis		Z-score ^ξ	Range [£]		Corrected Item-Total Correlation
		Mean	SD	SEsk			SEk			Min	Max	
Q1	132	4.31	0.97	-0.66	0.21	-3.14*	-0.07	0.42	-0.17	2	6	0.85
Q2	132	4.64	0.80	-0.59	0.21	-2.81*	-0.09	0.42	-0.21	3	6	0.87
Q3	132	4.56	0.82	-1.03	0.21	-4.90*	1.49	0.42	3.55*	2	6	0.83
Q4	132	4.58	0.83	-0.80	0.21	-3.81*	0.60	0.42	1.43	2	6	0.98
Q5	132	4.25	1.01	-0.79	0.21	-3.76*	0.24	0.42	0.57	1	6	0.83
Q6	132	4.11	0.99	-0.81	0.21	-3.86*	0.97	0.42	2.31*	1	6	0.80
Q7	132	4.75	0.76	-0.63	0.21	-3.00*	0.36	0.42	0.86	3	6	0.82
Q8	132	4.71	0.83	-0.83	0.21	-3.95*	0.63	0.42	1.50	2	6	0.84
Q9	132	4.83	0.79	-0.92	0.21	-4.38*	1.32	0.42	3.14*	2	6	0.82
Q10	132	4.51	0.93	-0.84	0.21	-4.00*	0.48	0.42	1.14	2	6	0.82
Q11	132	4.75	0.97	-1.02	0.21	-4.86*	0.99	0.42	2.36*	2	6	0.82

* 44 surgeons used the OCAT in 132 observations of 79 residents from Orthopedics, General Surgery, or Obstetrics/Gynecology at the University of Ottawa.

φ Surgeons rated residents' performance on each item using a six-point scale ranging from 1= "I had to do" to 6 = "Functioning at level higher than a generalist in my specialty."

£ Range of actual ratings for each item across the assessments completed for 132 observations. Min indicates minimum; Max, maximum.

ξ Z-score calculations based on the following formulas: Skew/SEskew = Z-score, if $x > \pm 1.96$ then significant; Kurtosis/SEkurtosis = Z-score, if $x > \pm 1.96$ then significant.

Note * $p < 0.05$

Table 15 also displays the item-total correlations associated with the items one to eleven. As shown in the table, the item-total correlations are relatively high ranging from 0.80 to 0.87 indicating that ratings on the items were similar. These ratings are explored in more detail in

Table 16, which displays inter-item correlations (Spearman's rho). The correlations range from 0.59 to 0.85 with ratings on items 2, 3, and 4 being highly correlated with one another¹.

Table 16: Inter-Item Correlation Matrix

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Q1	1.00	0.76*	0.68*	0.72	0.75*	0.72*	0.67*	0.66*	0.62*	0.66*	0.61*
Q2	0.76*	1.000	0.80*	0.84*	0.71*	0.69*	0.66*	0.78*	0.71*	0.67*	0.70*
Q3	0.68*	0.80*	1.000	0.78*	0.68*	0.63*	0.65*	0.74*	0.67*	0.66*	0.71*
Q4	0.72*	0.84*	0.78*	1.00	0.76*	0.71*	0.73*	0.77*	0.67*	0.70*	0.73*
Q5	0.75*	0.71*	0.68*	0.76*	1.00	0.75*	0.62*	0.64*	0.58*	0.61*	0.59*
Q6	0.72*	0.69*	0.63*	0.71*	0.75*	1.00	0.64*	0.68*	0.58*	0.65*	0.56*
Q7	0.69*	0.66*	0.65*	0.73*	0.62*	0.64*	1.00	0.77*	0.76*	0.63*	0.71*
Q8	0.66*	0.78*	0.74*	0.77*	0.64*	0.68*	0.77*	1.00	0.79*	0.66*	0.70*
Q9	0.62*	0.71*	0.67*	0.67*	0.58*	0.58*	0.76*	0.79*	1.00	0.66*	0.78*
Q10	0.66*	0.67*	0.66*	0.70*	0.61*	0.65*	0.63*	0.66*	0.66*	1.00	0.76*
Q11	0.61	0.70*	0.71*	0.73*	0.59*	0.56*	0.71*	0.70*	0.78*	0.76*	1.00

* Correlation is significant at the 0.01 level (2-tailed).

Generalizability analysis. A generalizability analysis was used to determine the reliability of the scale. The model that was used in this analysis was mixed ANOVA with forms (f) nested within residents (r) nested within discipline (d) and crossed with items (i). Individual residents were the object of measurement. As shown in Table 17, which lists the facets and variance components for this model, 34% of the variance was accounted for by the participant residents within their disciplines (r:d, comparing OCAT scores for resident A in general surgery to OCAT scores for resident B in general surgery). Differences between individuals' forms (f:r:d) accounted for 33% of the variance (comparing OCAT 1 score for resident A in general surgery to OCAT 2 score for the same resident). Facets involving items (i, d:i, ri:d) accounted for

¹ A factor analysis, performed to confirm the underlying dimensions of the tool, showed all the items loading onto one factor. Given the small sample size, the formative nature of the tool, and the redundancy of this data when combined with high inter-item correlations and a small combination of the items to the variance (ANOVA), the factor analysis data has not been included here.

only small amounts of the variability in scores which confirms the ratings are highly correlated (see Table 16). Other factors contributed 25%. Table 18 gives more of an explanation of all the facets included in the model. The model was constrained by the fact only one rater observed each interaction and inter-rater reliability could not be calculated for the same reason.

Table 17: Ottawa Clinic Assessment Tool Variance Components

Facet*	Variance Component	% Variance
Discipline (Ortho, Gen Surg, Obs/Gyn) (d)	0.00	0
Residents within Discipline (r:d)	0.29	34
Variance within Resident; Forms within Discipline within Participants (f:r:d)	0.28	33
Items (i)	0.05	6
Discipline within Items (d:i)	0.01	1
Participant Items within Discipline (ri:d)	0.01	1
Everything else (fi:r:d)	0.21	25

* Facets: d indicates discipline (Ortho, Gen Surg, Obs/Gyn); r, residents; f, forms; i, items.

Table 18: Variance Components Explained

Facet	Explanation
f:r:d	The variance attributable to the individual OCAT forms for individual residents within their discipline i.e. comparing OCAT A for resident A within general surgery to OCAT B for resident A within general surgery (potentially attributed to contextual or rater factors).
r:d	The variance attributable to the residents within their disciplines, i.e. comparing resident A within general surgery to resident B within general surgery.
d	The variance attributable to the disciplines of origin of the resident being assessed (orthopedic surgery, general surgery, obstetrics & gynecology).
i	The variance attributable to the items on the tool.
d:i	The variance attributable to the discipline for the individual items. For example, did item ratings in orthopedic surgery vary from those in general surgery.
ri:d	The variance attributable to the individual resident participants within their disciplines looking at individual items. For example, did item ratings for resident A within general surgery vary from item ratings for resident B in general surgery
fi:r:d	The variance attributable to the form, items, participants, their disciplines, plus random error. In ANOVA terms, this is the residual.

The variance components displayed in Table 17, were used to determine the reliability of the scale. The formula for calculating the reliability is located in Figure 6 below. For an average of three forms submitted per person the resulting g-coefficient was 0.74. For four forms it increased to 0.79, and with five to 0.83.

Figure 6: Generalizability Equation

The following generalizability equation was used:
$$G = \frac{\sigma_{r:d}^2 + \frac{\sigma_{ri:d}^2}{ni}}{\sigma_{r:d}^2 + \frac{\sigma_{ri:d}^2}{ni} + \frac{\sigma_{f:r:d}^2}{nf} + \frac{\sigma_{fi:r:d}^2}{nfn}}$$

Subscale question 12 (Q12) & 13 (Q13) (technical skills). Q12 was completed on 49 of the forms, and Q13 on 55. The scores for item 12 and 13 were skewed and item 13 also showed kurtosis in its distribution of scores (see Appendix R for histograms of distribution for items 12 and 13). A range of choices between 3 and 6 were chosen for both items. See Table 19 for the complete descriptive statistics of the subscale.

Table 19: Descriptive Statistics of Items 12 & Q13 (Subscale)

OCAT	N	Rating ^φ		Skewness		Z-	Kurtosis		Z-	Range [£]		Corrected
Item						score ^ξ			score ^ξ			Item-Total
		Mean	SD		SEsk			SEk		Min	Max	Correlation
Q12	49	4.65	0.69	-0.97	0.34	-2.85*	0.77	0.67	1.15	3	6	0.57
Q13	55	4.91	0.65	-0.77	0.32	-2.41*	1.81	0.63	2.87*	3	6	0.57

φ Surgeons rated residents' performance on each item using a six-point scale ranging from 1= "I had to do" to 6 = "Functioning at level higher than a generalist in my specialty."

£ Range of actual ratings for each item across the assessments completed for 132 observations. Min indicates minimum; Max, maximum.

ξ Z-score calculations based on the following formulas: Skew/SEskew = Z-score, if $x > \pm 1.96$ then significant; Kurtosis/SEkurtosis = Z-score, if $x > \pm 1.96$ then significant.

Note * $p < 0.05$

Spearman r correlations between the subscale totals (items 12 & 13) and the scale totals (items 1-11) reached significance ($p < 0.02$, 2-tailed). See Table 19 for descriptive statistics unique to each item and to the combined subscale, and Table 20 for the correlation data.

Table 20: Technical Skills Subscale Q12 & Q13 Correlations

		Correlations			
		Q12	Q13	Sub Scale Total	Scale Total
Q12	Spearman r Correlation	1.00	0.49**	0.87**	0.63**
	Sig. (2-tailed)		0.00	0.00	0.00
	N	49	49	49	49
Q13	Spearman r Correlation	0.49**	1.00	0.84**	0.69**
	Sig. (2-tailed)	0.00		0.00	0.00
	N	49	55	55	55
Sub Scale Total	Spearman r Correlation	0.87**	0.84**	1.00	0.74**
	Sig. (2-tailed)	0.00	0.00		0.00
	N	49	55	55	55
Scale Total	Spearman r Correlation	0.63**	0.69**	0.74**	1.00
	Sig. (2-tailed)	0.00	0.00	0.00	
	N	49	55	55	132

** Correlation is significant at the 0.01 level (2-tailed)

Global assessment item. Residents receiving the answer of “no” to the global assessment item had a mean score on the remainder of the items on the OCAT of 3.79, and those receiving a “yes” had a mean score of 4.88. The main effect is significant ($F(1,121) = 89.41$, $p < 0.001$) with a large effect size of 0.425. See Table 21 for descriptive statistics of the relation of the global assessment item to the remainder of the items.

Table 21: Global Assessment Item

Answer	Mean Scale Total	Std. Dev	# of Forms
N	3.79	0.79	39
Y	4.88	0.48	84
Total	4.54	0.78	123

Other analyses.

Discipline. As shown in the g-study, discipline by itself did not account for a large amount of the variability in scores. The point of this analysis was to explore this factor in more detail. Ratings were averaged across items (Q1-11) to produce a mean scale total. There was a non-significant main effect of discipline ($F(2,129) = 2.62$, $p=0.08$), effect size 0.04. Since the differences between discipline approached significance, a post hoc test was completed and showed that the mean rating for obstetrics and gynecology was higher than the mean from rating for orthopedic surgery ($p=0.024$), but there were no other differences between the other two specialties ($p>0.05$). For the technical skills subscale, the main effect of discipline was not significant ($F(2,52)=2.608$, $p=.08$), effect size of 0.09. Due to the low number of forms completed for items 12 and 13, no post hoc testing was completed. See Table 22 for the descriptive statistics related to discipline for items 1-11, and Table 23 for the technical skills subscale.

Table 22: Effect of Discipline Q1-Q11

Discipline	Mean Scale Total	Std. Dev	# of Forms
Gen Surg	4.54	0.64	65
Obs Gyn	4.86	0.42	21
Ortho	4.41	0.98	46
Total	4.55	0.76	132

Table 23: Effect of Discipline Q12-Q13 Subscale

Discipline	Mean Scale Total	Std. Dev	# of Forms
Gen Surg	4.55	0.71	19
Obs Gyn	4.81	0.51	21
Ortho	5.00	0.46	15
Total	4.77	0.59	55

Effect of PGY level. There was a significant effect ($F(3,128)=32.02$, $p<0.001$) of PGY level with a large effect size of 0.43. The groupings were as follows: PGY1/2 (junior residents) grouped, PGY3, PGY4, and PGY5. Total item score of items 1 to 11 (not including the subscale) was the dependent measure. The total clinic score differed significantly ($p<0.001$) between all levels, except PGY4 and PGY5 with a large mean effect size of 0.43. Post hoc tests revealed that PGY 1 and 2 were lower than 3 ($p=.03$), lower than 4 ($p<.001$) and lower than 5 ($p<.001$). PGY 3 was lower than 4 ($p<.001$) and 5 ($p<.001$). PGY 4 was not lower than PGY 5 ($p=0.12$). Mean scores per resident were as follows: PGY1/2 3.71, PGY3 4.08, PGY4 4.82, PGY5 = 5.02 (Table 24).

Of note, only 19 forms were submitted by junior residents (PGY1 and 2). The majority of forms were submitted by PGY4 and above ($N= 113$).

Table 24: Effect of PGY Level Q1-Q11

Year of Residency	Mean Scale Total	Std. Dev	# of Forms
PGY 1 & 2	3.71	0.87	19
PGY 3	4.08	0.70	32
PGY 4	4.82	0.42	40
PGY 5	5.02	0.43	41
Total	4.55	0.76	132

For the subscale (items 12 and 13), the groupings were the same as above. The subscale clinic scores showed a difference as well ($F(3,51) = 4.95$, $p=0.004$) with a moderate effect size of 0.23.

Post hoc tests revealed that PGY3 differs from PGY4 ($p=0.007$) and PGY5 ($p<0.001$). These results are based on a low n value and require further testing in a full study. See Table 25 for descriptive statistics related to the subscale broken into PGY levels.

Table 25: Effect of PGY Level Q12-Q13 Subscale

Year of Residency	Mean Scale Total	Std. Dev	# of Forms
PGY 1 & 2	4.25	0.35	2
PGY 3	4.22	0.71	9
PGY 4	4.83	0.59	18
PGY 5	4.96	0.42	26
Total	4.77	0.59	55

Clinic type: specialist vs. generalist. Another question was if there was an effect of which clinic the residents were being assessed in (specialist versus generalist clinic). For this reason the correlation between scale total scores and the type of clinic was assessed. There was a significant main effect of clinic type ($F(1,130) = 4.096, p=0.045$), with a small effect size of 0.031 (Table 26).

For the technical skills subscale, the main effect was not significant ($F(1,53) = 0.046, p=0.83$) with a very small effect size of 0.001 based on a small n value (N generalist = 20 forms, N specialist = 35). See Table 27.

Table 26: Effect of Clinic Type Q1 – Q11

Clinic Type	Mean	Std. Deviation	# of Forms
Generalist	4.41	0.86	63
Specialist	4.67	0.64	69
Total	4.55	0.76	132

Table 27: Effect of Clinic Type Q12 – Q13 Subscale

Clinic Type	Mean	Std. Deviation	# of Forms
Generalist	4.75	0.70	20
Specialist	4.79	0.53	35
Total	4.77	0.59	55

Short answer question analysis. The short answer questions asked the staff raters to name one specific thing the resident did well (Q14), and one specific thing they could improve (Q15). These were coded the primary researcher and the research assistant (C.S.) independently to quantify how many of the answers directly correlated to items on the OCAT, and how many were novel items of feedback.

Question 14 – Specific aspect of clinic done well. Question 14 was completed 124 times out of 132 forms (94% of the time), with 14 left blank. 63 of the 124 responses to Q14 were general supportive statements (e.g. good job, great clinic, etc.). Only one comment under question 14 was *not* positive. The most common directly referenced item from the tool in short answer question 14 was Q2 (history). Second and third most frequently correlated items were Q6 (management plan) and Q4 (case presentation). Only five comments under Q14 were coded as “very specific elements of clinic” that did not directly relate back to the items already on the OCAT (see Table 28).

Table 28: Comments Relating to Specific Elements of Clinic Done Well

Excellent knowledge of metastatic disease management.
Confidence with tackling difficult clinical tumour cases.
Aware of patient anxiety and responded appropriately.
Work-up of confused patient with chronic pain.
Very compassionate to patients, good biopsy technique.

Question 15 – Specific suggestion for improvement. Question 15, the location on the tool providing an opportunity for constructive criticism, was more often left blank (27% of the time). It was completed 96 times out of 132 forms, with 36 not filled in. An additional 13% of responses (10 of 132) were “None” or “Nothing”, implying there was nothing to improve. Therefore 40% of the time, no constructive criticism of any kind was offered.

There were some responses within Q15 that directly correlated to items 1 to 11 on the OCAT (see Table 29). The most frequently mentioned item for improvement was Q6 (management plan), with Q12 (technical skills), Q8 (documentation), and Q9 (collaboration) gathering a few mentions.

Another coding option was added for Q15. It was observed during an initial read-through of the results that a frequent piece of constructive criticism given was “read more”. This statement occurred seven times (8.1% of comments made). 31 forms (36% of comments made) contained very general comments (e.g. keep increasing knowledge base, keep at it, etc.). Another interesting observation was the prevalence of a positive comment (or a veiled compliment) being given in place of an item to be improved. This was a separate code as well which occurred 15 times (17% of comments made).

Table 29: Short Answer Questions –
How Many Comments Correlated Directly to Items on the OCAT?

	Item on OCAT	Did Well (Q14)	Could Improve (Q15)
Q1	Clinical Preparedness	2	2
Q2	History	26	2
Q3	Physical Exam	10	2
Q4	Case Presentation	15	1
Q5	Differential Diagnosis	2	4
Q6	Management Plan	17	16
Q7	Patient/Family Communication	11	4
Q8	Documentation	10	5
Q9	Collaboration	1	1
Q10	Time Management	2	5
Q11	Attitude	7	0
Q12	Technical Skills	13	7
Q13	Situational Awareness	0	0
Total		116	49

Step 6B: Feasibility check—length, clarity, and usability.

Informal feedback throughout the pilot project. Informal feedback was collected verbally and by email from study participants throughout the pilot study. Two significant contributions were made to the project by this means. Other informal feedback helped with the writing of feedback survey questions.

An obstetrics and gynecology staff observed verbally that the orientation of the yes/no answer to the global assessment item was in reverse order from left to right on the page (yes first, no second), whereas the anchor scale was in ascending order from left to right (one to five increasing). He stated that this was initially confusing to him, and each time he filled out the

form he had to stop and consider yes/no carefully. At the time of the study, the decision was made to keep the order the same due to the size of the change, but this decision will be re-evaluated for future studies. Previous work has demonstrated that reversing the scale answer usually leads to more errors, both by the raters and the researchers (Streiner & Norman, 2008, p.52).

A staff rater from general surgery pointed out that he did not realize that the distinction between scale items five and six meant the distinction between a competent graduate and fellow level. The implication that staff raters were not reading the instructions at the top of the form carefully enough was considered a piece of response process evidence. It was decided that this was a serious enough issue for a change to be made to the form halfway through the pilot study. In early October, a vertical line was placed on the anchor scale between five and six, and a flag stating “at fellow level” hovered over the six column. Staff raters were pleased with this change.

Feedback sessions. Results from the feedback sessions are presented below in question/answer appearance format for clarity. These were the questions that were posed to the feedback groups.

Staff feedback session.

What did you think of item 1 – clinic preparedness?

Staff agreed that this was a difficult item to complete. The phrasing of the item did not match up with the anchor scale. As well, individuals were interpreting the item differently. Some considered clinic preparedness to be reading and knowledge base before coming to clinic; others did not grade residents well unless they had specifically looked up the patient list before the clinic. One of the surgeons commented that preparedness comes into other items on the tool (time management, differential diagnosis, and management plan) and if it were removed, it would not be missed.

What did you think of item 11 – attitude?

This item was also difficult to complete and had a variety of interpretations, including: engagement, professional behavior, and attitude towards being present in clinic. The overarching comment, however, was that this is more of a yes/no question. It is difficult to rate a resident’s attitude on a five-point entrustability scale.

When working in a sub-specialized clinic (ex. uro-gynecology, hepatobiliary, spine, etc) how did you use the rating “5”?

There were some comments from sub-specialized staff surgeons throughout the pilot project that they were having trouble understanding the frame of mind they should be using while assessing with the OCAT. Some were incorrectly assessing the residents’ ability to run a *sub-specialty* clinic independently. While others were using the OCAT as it was intended. Whenever this frame of reference problem was encountered, the primary researcher would try to re-orient them by asking them to imagine that they were assessing a resident who was completing the final year oral exam.

Interestingly, in the feedback session, the same frame of reference difficulty was not noted. Staff surgeons said that making the specialist/generalist distinction was not challenging for them. There was a comment that all academic staff surgeons need to be aware of what a generalist surgeon needs to know anyway if they are going to teach.

However the residents had a different perspective in their feedback session. They were concerned that some of the staff were not reading the pre-amble instructions on the tool and therefore did not understand how to use the anchor scale and the global assessment item. This concern was voiced by three of the six residents. The other three residents did not express concern over their staff using the tool incorrectly in sub-specialized clinics.

Did you struggle with the yes/no (global assessment item) question? Why or why not?

Comments were very positive regarding the nature of the global assessment item. Users felt that it made them, “Take a side” and initiated more constructive criticism than would otherwise have been given. Overall, they understood its usefulness for a competency-based program, and agreed that by PGY-5/last year of residency, there is great concern if residents are not receiving a “Yes” on the global assessment item. Two program directors stated this would be very useful for their program of assessment.

The main concern with the item was the unclear nature of the word “manage”. Two out of five surgeons felt that it was confusing when working with junior residents. A PGY1 who is able to keep seeing patients independently, present an organized case, and come ask for help when needed is “managing” a clinic well in their opinion. Another staff made the suggestion to use the synonym “run” a clinic, instead of “manage”, as it is more specific to a staff clinician. A final comment was offered from another staff that it was helpful for her to think of this question as asking, “If you weren’t here today, what would’ve happened to your patients?” If she felt

confident they would have been well cared for, then she gave a “yes”. If not, the resident received a “no”.

Again, it was interesting that the residents did not entirely share the staffs’ opinion. One resident (the least advanced, PGY1) said, “The first time I had this form filled out I got all 4’s and 5’s and a ‘yes – you’re safe to independently run clinic’. I’m an R1, and I assure you, I’m not that stellar”. Two out of four others agreed that this occurred when they had forms filled out as well. A PGY3 commented:

A lot of times staff would make comments like, ‘I think you are where you should be for your level of training.’ That may indicate that they are using this scale as they envision in their head, and not using the scale as explained on the form.

What did you think of the rating “6”? How did you use it?

Early in the study, the research team decided to use a six-point scale so Fellowship training programs could potentially use the OCAT in the future. However, staff raters did not comment on potential usefulness for this purpose. They did state that they used the six very similarly to how they would have used a five if the scale had been only a five-point scale (albeit sparingly).

One surgeon felt that the six provided the opportunity to recognize excellence in senior residents. Given that CBME requires competence, while encouraging excellence, this is an important point to consider. There were two comments that agreed on the preceding point. A resident receiving a six is able to customize a surgical plan to the individual. “They don’t just know the book answer, they’re thinking of how the book answer affects the person sitting in front of them.” [Staff Surgeon]

Originally the research team had envisioned the six-point scale also helping guide the raters away from correlating performance directly to one of the five PGY years (norm referencing). Study participants made no comments to this effect.

Overall, was the OCAT user friendly or not? Why?

In general staff surgeons thought the items were excellently worded, relevant to surgical clinic, and not onerous to use. They unanimously agreed, however, that the OCAT was too long (it took up a legal size piece of paper). If possible, they asked for the print to be bigger so it is more legible. The study mechanism made it very easy to complete, since the forms were available in clinic and they could just transfer them between study envelopes.

One of the surgeons also commented on the \$10 gift card given to residents who completed forms during the study. He believes that this extra reward was the motivation behind any OCATs being completed. There was concern that after study completion, with no reward being offered, the number of OCATs turned in would decrease.

How feasible is it to do an OCAT at the end of the clinic?

For the staff surgeons, it was feasible. They agreed that it took approximately five to ten minutes to complete. However, three of the five surgeons stated that they did not think this was enough time at the end of clinic to give appropriate feedback. There was also a generalized concern that it's easy and quick to give excellent residents feedback, but it's more challenging for struggling residents. There was unanimous agreement that the OCAT is more feasible for them to use at the end of clinic than a direct observation tool.

There was near unanimous agreement on feasibility in the resident group as well. It was the resident's responsibility to bring the OCAT to their supervisor at the end of clinic for completion. Given this fact, one resident noted that,

It doesn't sink in so well with the quick flow of clinic. If at the end of a busy day of clinic they [staff surgeon] seemed not in the mood to fill one out, I sometimes wouldn't ask. I had to pick the right day and be aware.

In response, another resident agreed but qualified by saying, "Most of the time we were about to run off, but as long as I remembered to ask, we were able to fit it in".

One of the program directors summarized the feeling of both groups well when he said: The reality is we're going to see more and more of these forms [assessment tools] because the college will require them and we'll be having to figure them out. I think that this is actually quite a good form compared with others... It is time-consuming, but it isn't arduous. [Program Director]

Did completing an OCAT impact on your verbal feedback to the resident?

This was the most striking result of the OCAT pilot project. Residents and staff surgeons unanimously agreed during feedback sessions and on the survey that the OCAT increased daily formative feedback. One resident added to this during the resident feedback session:

It impacted, in that it [the feedback] was more constructive. Often I'll just get [from the staff surgeon], 'That was a good clinic, see you tomorrow' or 'Hey let's grab a coffee and chat', and then you don't chat about anything formal. [General Surgery resident]

Staff surgeons also affirmed that the OCAT increased formative feedback opportunity. They noted that it may not be the tool itself, but the window of time it created by sitting down to complete the form. A staff surgeon who used the tool with multiple residents said, “I think I gave them more feedback when I was using the tool just because it created a window where we had to sit and talk.” It was also noted that the items on the OCAT focused the feedback given, and broke it down into more specific information than the staff would normally have commented on. One surgeon put it this way:

I usually spend most of my [feedback] time teaching hard clinical information regarding the pathology. For the rest of the stuff I would usually gestalt the rest. This form helped me break down the rest of that information – attitude, etc – for the first time.

[General Surgery staff member]

Is there anything you would change about the OCAT?

There was one further suggestion made to change item ten “time management” to more clearly represent time management of an *entire* clinic. This staff surgeon thought it was a critical skill for residents to develop competency in, and did not want it confused for time management with individual patients.

Would you prefer to do the OCAT on paper or electronically on your iPad? Why?

Responses to this question in the staff and resident groups were opposite of what the study team expected. Staff surgeons said they would definitely rather have electronic forms. Residents stated that paper was easier because electronic forms often require three extra steps (passwords, firewalls, etc.) before you can fill them out. Residents expressed concern that if forms were not as readily available as they were during the study (hanging in an envelope in front of the clinic desk), staff would not take the time to wait for them to load an electronic form, and/or the resident themselves would forget to ask. One resident was also concerned that an electronic form may cause less verbal feedback if the staff is occupied with typing comments into the iPad instead of just talking to him. Both groups did comment that if the OCAT was available electronically, important features for it to have would be: easy access (an App would be best), available for iPad not just PC for portability, and it would likely need to be mandatory (repercussions if you didn’t complete it).

How many times could you imagine using the OCAT for resident assessment throughout the year?

Staff and resident groups both commented on a redundancy that developed if the same staff completed the OCAT for the same resident more than once per week. A PGY-3 General Surgery resident said, “By the time I got two of them [OCATs] from each of the two surgeons, I brought it to them the next time and they said, ‘I just did this yesterday, and you did the same today’”. This redundancy was seen more in the application of the anchor scale to the items, than to the short answer feedback.

The consensus in both groups was that individual residency programs would have to decide on an appropriate number based on a few factors: how many types of clinic occur during that rotation (e.g. cancer clinic is very different from post-op clinic), how many weeks is the resident on service, how many surgeons work on that service, and how often the resident is actually able to attend clinic (rather than the operating room).

Resident feedback session.

Did having a staff complete an OCAT on your performance have any impact on the verbal feedback you received in clinic?

See staff question regarding feedback above.

How often did staff review the OCAT with you? How did that occur?

Although recognizing that completion of the OCAT was a time-added task, none of the residents felt any resistance from supervisors to completing the form when they did ask. Completion of the OCAT always occurred at the end of clinic, unless the resident had to leave early. Usually the staff completed the OCAT with the resident present, but two of the feedback session participants had heard of other residents about staff that took the forms to their office to fill out independently.

How did you feel when you received a score of three or lower on an item?

Residents were eager to talk about this issue, but did not come to a conclusion. Five out of six agreed that the anchor scale made lower scores more acceptable, and easier to take, than a scale using “below average”. A PGY1 said, “I don’t find a one or two offensive because the scale is not saying I’m totally incompetent, only that I needed help”. One resident (a PGY3) commented that he had received some three’s and he had “no problem with them”. As he said

this, he was looking back at the OCATs he received and noted, “These three’s were completely appropriate. They [the staff surgeon] are right, they did need to walk me through that part of the management. It’s framed in the way that ‘you need help’, and I’m okay with that.” One resident felt that he would never be happy with a one or two, but that it would definitely teach him something, “It’s not good [receiving a one or two]. But if it’s honest feedback, and you have time for constructive conversation about how to improve, then it’s valuable.”

However, residents also noticed that they rarely scored a three or lower. There was actually concern over this and the majority of residents in this group said they want more constructive criticism. A PGY4 General Surgery said,

I always think that it’s better when someone actually points out a way to be better. I find the ones where someone says, “Oh, you’re a five”. And I’m like, “really, I know that’s not the case”. I know they don’t feel that way and I think, “why don’t you just write down how you really feel and we can make this useful”. [PGY4 General Surgery]

Following this there was a discussion of how high performance athletes don’t improve by just hearing good feedback all the time. Resulting in a PGY3 General Surgery saying loudly, “I just don’t know how anyone gets better, if they’re great all the time”.

Any other negatives of using the OCAT from your perspective?

A few negatives of WBA surfaced that were not specific to the OCAT form. Residents thought staff surgeons were not always being completely honest in an effort to avoid conflict. A resident commented on the opposite effect, how uncomfortable it can be to work with a staff for a whole month if they give you negative feedback on the first day. Otherwise, they had no specific negatives to add than what had already been mentioned.

Any other positives of using the OCAT from your perspective?

Residents surprisingly commented on the bright green color of the form, saying it reminded them to fill it out. They also re-iterated the OCAT’s ability to point out areas of deficiency. A particular benefit perceived by a PGY4 ortho resident was the timeliness of the assessment. “Having the context of the clinic helps because it allows for examples to be given. They buy it more because there are “discrete examples,” she said.

Residents were very positive about the effects of the study on their residency programs. One orthopedic surgery resident summarized her feelings this way:

The tool served a couple of functions. One was just to remind me that as a learner I need to ask for feedback. As a learner, I get caught up in responsibilities and duties and don't get around to that. Also, it was real time. Rather than being three weeks down the line when I'm done and can't respond to it. If they give me something to work on then I can show them I'm trying to improve it. [Orthopedic Surgery PGY2]

Residents believed that the effects of the OCAT extended beyond the clinic environment as well. "It [the OCAT study] built an awareness on both sides that evaluation is important. This makes it happen more informally as well," said a PGY3 General Surgery resident.

Section 5: Discussion

Introduction

Medical educators working to bring CBME to fruition in practice have found it difficult to balance assessment priorities. On one hand, assessment should be objectives-focused, measuring discrete components of an individual resident's performance in as unbiased a manner as possible. On the other hand, the social nature of a resident's work makes it difficult to separate their performance from the rest of the team, and objective attempts to assess outside of the workplace become less representative of actual in-training performance (Ginsberg et al, 2010). The question of what role subjective, context-based assessments should play in this new competency-based workplace is being raised (Govaerts & van der Vleuten, 2013). The gestalt of a surgeon who is asked if the resident they are working with is competent means something, but the best way to capture this gut feeling is not obvious.

Even within this milieu of the rising role of subjectivity in WBA (Hodges, 2013), it is still crucial to demonstrate the validity evidence behind assessment tools (including psychometric data). The goal of this project was to develop a workplace based assessment tool that surgical supervisors can use to assess the progression of their residents towards competence. Strong influences on the development of this tool were feasibility concerns and increasing evidence that WBA tool components such as construct-aligned scales can improve tool performance (Crossley et al., 2011). At a high level, the goal of the OCAT is to contribute to a CBME assessment program by improving daily formative feedback in the clinic. Evidence to demonstrate that this occurred during the pilot project as well as validity evidence for its performance as a psychometric tool is outlined here.

Content Evidence

Pertaining to purpose. As a first step in determination of content, it is important to consider the purpose of an assessment tool. Plans for this clinic assessment tool began when team members at the Ottawa Hospital observed that only end-of-rotation ITERs were being completed for non-technical skill assessment, and they were not based on daily documented assessments. As well, many residents were commenting that they rarely received feedback from their staff supervisors on their non-technical skills. The feedback that was given often did not appear on their ITERs in meaningful, detailed form as seen in other studies (Dudek et al., 2008).

To help staff surgeons pass on any daily formative feedback already occurring, and to stress the key place of timely feedback in a CBME system, the primary researcher hoped to identify (or develop if necessary) a feedback tool succinct and relevant enough that surgeons would actually use it in a busy surgical context. Noting as well that since staff surgeons spend a lot of one-on-one time with their residents in clinic, potentially this could be a target environment for a low-stakes assessment opportunity and well as the teaching moment key to formative feedback (Van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010).

Formative assessments (assessments which initiate feedback) are used to guide improvements, whereas summative assessments help decision makers make crucial promotional choices (Guskey, 2014). Key features of a valid *formative* workplace based assessment program can be highlighted using these adjectives: continuous, frequent, over-time (not one point-in-time only), and requiring feedback (Touchie, 2013). The true goal of formative assessment is giving feedback to residents to notify them of areas for improvement on an ongoing basis (Holmboe et al., 2010), influenced by the context (case difficulty, patient personality, amount of sleep). Sampling from multiple contexts is crucial for formative assessment and includes performance under formal test conditions and in unobtrusive clinical situations.

Therefore, if the purpose of creating this tool was to increase and efficiently document feedback given in surgical clinic, developing a formative assessment tool made much more sense than a summative one. This plan was made with the acknowledgment that program directors will likely use OCATs as evidence that a trainee has met a set standard (a summative role). However there was no plan made to set a pass score, as would have been necessary for a purely summative tool, and the items were written in such a way that they would hopefully enhance opportunity for rich feedback. A further discussion of whether the tool ultimately developed (the OCAT) was successful in stimulating efficient formative feedback can be seen later.

Pertaining to content. The goal of an assessment tool is to measure a *construct* or multiple constructs, the intangible and abstract concepts that can be inferred from behavior. For example, educational achievement is a construct often represented by scores on multiple-choice exams in school. A starting point for constructing a tool is to investigate previous definitions and explanations of the construct to help with filling in test content. The construct of the OCAT is surgery resident independent clinic competence over a day of clinic and so a literature search of previous definitions of surgical clinic competence was undertaken. Although a lot of

discussion has occurred to define surgical competence and how the non-technical skills embodied in the CanMEDS roles can be blended (Fontes, Selden, & Byrne, 2014; Sidhu et al., 2004), no clear definition of what components are essential to manage specifically a surgical clinic could be found.

After a thorough search showed this gap in the current literature, the process began to gather content to be included in the clinic competence assessment tool. The CanMEDS competencies were a potential place to start, especially since surgical program directors need to give an account of how they are assessing their residents on these roles at accreditation time. However, leading up to this study, it seemed that there might be an interpretation issue occurring in which surgeons and surgical residents do not find the CanMEDS terminology applicable to their daily work environment. For example, when asked on an assessment form to indicate a rating for their residents' advocacy skills, surgeons anecdotally do not know what that means. Many authors have written about this, expressing that there is a tension between the competency framework and faculty's experience in the daily assessment of residents (Lurie, Mooney, & Lyness, 2009; Zibrowski et al., 2009). "This means that faculty members have to try to fit the round peg of their observations and judgments in to the square holes of an evaluation form, often resulting in meaningless ratings and comments entered only for the purpose of getting the evaluation done" (Hanson et al., 2013, p.3). As well, disconnect has been noted between the original holistic intent of the CanMEDS roles and what is assessed in little categories on a daily basis, causing misunderstanding during assessment (Ginsberg et al., 2011).

Therefore, instead of constructing the tool using existing CanMEDS terminology, consensus groups of surgical experts were gathered to develop test items that would use commonly accepted language in the surgical community and be as construct-aligned as possible. It was hypothesized that following the study, tool items could be 'translated' back to their CanMEDS roots (using the 'spirit' of the CanMEDS roles) as an aid for program directors if necessary. These two consensus groups, as well as the national expert survey, were designed to come to content expert agreement on what surgical clinic competence is.

The nominal group technique was chosen to allow for an item-generating process limiting individual biases (Campbell & Cantrill, 2001). As much as possible, all surgical divisions in the department at the Ottawa Hospital were included in these discussions, at both the resident and staff level, since there are performance elements in clinic that likely occur in some specialties more frequently than others. For example, the first consensus group discussed whether or not to

include technical skills in the tool and determined that for them these skills were sufficiently assessed in other environments. However, in the second consensus group, there was a surgeon who frequently runs procedure-based clinics and thought that there are unique enough elements to technical skills performed in the clinic (awake patient, situational awareness) that they should be assessed on the OCAT. As a direct result of this discussion two items on the pilot project OCAT refer to technical skill performance (one of these refers to non-technical aspects of procedural skills). Sampling widely in terms of surgical division representation was crucial to make the tool content acceptable to all surgeon-users when complete.

Following the local consensus group process, a wider geographic sample was taken. Key elements of clinic performance in Ottawa may be subject to bias just due to local training culture. When trying to develop a nationally applicable assessment tool, a national sample should be taken (Humphrey-Murto et al., *manuscript in preparation*). The national survey added to the content evidence of the consensus group data, since the majority of responses from the surgical education experts surveyed lined up with the items generated locally. No entirely new content items came up in the external survey. Also, the high survey response rate (83%) shows a strong interest in this topic overall and affirms the relevance of the tool under development.

To further add to content validity, the items were edited for clarity. To remove all ambiguity both the primary researcher, the clinical study team members, and a two-week pre-pilot trial ‘tried out’ a preliminary version of the OCAT. Descriptors of the items were written by primary researcher with input from the study team to reflect discussions at the consensus group meetings within one week after the second consensus group finished. The final product, a 13-item list of elements contributing to surgical clinic competence, was then converted into the one-page assessment tool (including brief item descriptors) that surgeons who pre-piloted the tool thought accurately represented the consensus group discussions. After two weeks of pre-piloting of the tool, large relevance and wording issues had been confidently addressed and the OCAT was ready for full piloting.

During the feedback sessions following the pilot study, content validity issues relating to two of the items were addressed. Item one (clinic preparedness) and item eleven (attitude), were identified as controversial by the primary researcher based on informal discussions with users during the pilot study. Difficulty completing item one stemmed from its lack of concordance with the entrustability scale. It was problematic for raters to respond to *Clinic Preparedness* by making a rating somewhere between *I had to do* and *Independent*. As well, there was no

consensus between staff of what constitutes preparedness. They wondered if it was referring to reading before clinic, or to familiarity with the patient's clinical histories? Based on feedback session discussion, this item will likely be removed from subsequent iterations of the OCAT given that residents can demonstrate their preparation for clinic in other items (while describing their planned management of a patient for example). The surgeons who took part in the feedback session said the item could be removed and would not be missed, improving feasibility due to shorter length.

As well there was general disagreement over the meaning of the attitude item (item eleven). Staff surgeons definitely valued a resident who was able to take some ownership of the patients they interacted with. However, the complex concept of professionalism also came up in relation to attitude (both in the feedback sessions and on the national survey). Not only was it difficult to define the breadth of professional behavior (as has been demonstrated elsewhere) (Dall'Alba, 2009), but also the consensus was that this concept would better assessed with a dichotomous question (yes/no) rather than a rating scale. Either a resident was professional and had a good attitude, or they were not. Based on this feedback, this question will be turned into another yes/no question (with the goal of flagging professionalism issues) near the end of subsequent versions of the OCAT.

There are a few limitations in the manner content evidence was gathered. During consensus groups it is ideal to maintain anonymity during the list generation process to avoid intimidation by senior staff over junior staff and residents (Humphrey-Murto et al., *manuscript in preparation*). However a process using anonymous answer cards, or a silent ranking process, was unlikely to have stimulated the discussion the study team was hoping for. Therefore it is a possibility that junior staff and residents did not voice their true opinions. The post-study feedback sessions were separated into staff and resident groups to avoid this effect as much as possible.

Additionally, there was time constraint imposed by completing the consensus groups at the end of a workday (planned this way for recruiting reasons). Participants may have just agreed with the final list of items to avoid discussing them further, thereby causing a *false consensus* (Tammela, 2013). This is a limitation of all consensus groups and the national expert survey was performed to help mitigate the risk by checking for further content items that may have been missed.

Internal Structure Evidence

A psychometric analysis of the pilot study dataset was conducted to analyze the internal consistency and reliability of the OCAT scores. Three surgical divisions were recruited for this pilot project, not just to attain a wider sample, but also to ensure a sufficient number of assessment forms were collected within the study period. With 44 surgeon raters and 79 residents recruited (not all residents submitted forms), 132 complete OCAT forms were collected with an average of 3 per resident. The study period actually had to be extended to account for the slower accrual of forms than expected. Explanations for this slower accrual include: fewer residents attending clinic than expected due to competing ward or operating room duties (in general surgery especially), slow uptake of the tool in obstetrics and gynecology clinics as they have not been involved in previous similar research studies, residents rotating on blocks that did not include any clinic time at all (ex. acute care rotations in general surgery), and a low number of junior residents rotating through their home service during their PGY1 and PGY2 years (i.e. junior residents do a lot of off-service rotations).

65 of the 132 forms were collected from general surgery, likely due to a greater number of residents in the program and the primary researcher's personal connection with many of the residents. However, a large number (46) were also received from the division of orthopedics who had previously been involved with similar studies. Obstetrics and gynecology was recruited to the study with less marketing and without prior participation in similar assessment studies, and so they contributed the fewest forms. During this study, completion of the OCAT was not mandatory and some motivation for completion of forms came from a resident stipend distributed per form completed. In the future, if CBME programs are going to insist on having a certain number of WBAs completed, their completion will likely have to be mandatory to compete with other conflicting service and education priorities.

Item analysis. An item analysis was performed, with the two technical skills items (12 and 13) separated into a sub-scale since this section of the tool was only filled out if procedures were completed. A low number of OCATs included sub-scale data (item 12 was completed on 49 and item 13 on 55 of 132). The surgical divisions included in the pilot project have a low volume of procedure clinics, and these technical skill items will have to be tested in a division such as plastic surgery with more procedure clinics to fully elicit their merit.

The first eleven items had high mean ratings (greater than 4.0 out of 6.0 for all items). This is a worrisome effect as it could point to leniency biases occurring and failure to use the low end of the scale (Kogan et al., 2014; Gray, 1996). However on further analysis, it is clear that despite high mean ratings, staff surgeons did use a range of scores for each item. A rating of 1 or 2 was used on all items except for item two (history taking) and item seven (patient/family communication), which each had a minimum ranking of 3. Looking at the anchors of the entrustability scale (see Figure 4), a score of 1 on these items would mean the staff rater felt he or she had to completely do the task embodied in the item, and a 2 would mean the staff would have to talk a resident through it. Since history taking and communication skills are necessary skills to graduate from medical school, it is reasonable to expect that residents would achieve at least a score of 3 – “I had to direct them from time to time”, even as a PGY1. In contrast, item five (differential diagnosis), which showed a wide range of ratings assigned, is more of an advanced skill. It is entirely possible that junior residents may not have the ability to present a wide differential diagnosis and should be scoring 1’s and 2’s on the scale. One of the goals of an entrustability scale is to improve acceptance of lower ratings (both for staff and residents), therefore resident thoughts on their perception of lower scores were sought during the feedback sessions.

Interestingly, the range of scores for the technical skills subscale was from 3 to 6. Therefore all staff surgeons who scored their residents thought they only had to direct their residents from time to time. A likely explanation for this finding, which should be corroborated with future studies, is that staff surgeons were only scoring residents who physically performed procedures. They did not score those who observed a procedure as it was completed, or who were present in clinic that day but did not perform the procedures for another reason. In future studies, it will be important to encourage staff surgeons to use the rating score 1 and 2 if they demonstrated a procedure for a resident or if they felt uncomfortable with entrusting that task to them on that day. This is valuable feedback for the resident to receive.

Inter-item correlation data showed high correlations between the items (see Table 16) as has been shown for previous clinical performance scores in outpatient settings (Cook, Beckman, Mandrekar, & Pankratz, 2010; Wilson et al., 2014). A high internal consistency of 0.97 was seen. Potentially this high correlation could be a sign of halo effect, in which raters are unable to judge aspects of performance independently during one observation (Cooper, 1981). When considering these findings in isolation, they indicate that the OCAT can be significantly shortened since all items are all measuring one construct. However, although this is true for summative purposes, this tool was also designed to be formative. Given that all items were developed directly from the consensus group list of key elements of clinic competence, the value of an item for generating constructive feedback should be considered (Dudek et al., 2008). Future iterations of the OCAT will have to balance the number of items removed for the sake of conciseness and the formative benefit of keeping the items in place.

The technical skills items (12 and 13) also showed correlations with the remainder of the items, but due to the low number of responses on the sub-scale, more study is necessary to determine whether they should truly be analyzed as a sub-scale.

The global assessment item as a merged impression of performance. The global assessment item at the end of the OCAT may be another good reason to re-consider significantly decreasing the number of items. There is some evidence that one part of an assessment tool can influence how a rater completes another component (Gofton et al., 2012). The OCAT begins with 13 individual items all mapping to the same construct, and then continues on to a global assessment item asking for the rater's subjective impression of whether this resident is ready to independently run a generalist clinic. Analysis of the global assessment item data showed a significant difference between the mean scores of residents who received an answer of 'yes' to this statement and those who received a 'no' (mean scores of 4.88 and 3.70 respectively, see Table 21). Previous work points to the significant influence a checklist of items can have on the way a rater answers the global assessment item (Regehr, MacRae, Reznick, & Szalay, 1998).

It seems this list of items helps the rater develop a merged impression of the goal of the instrument, therefore the individual items plus the global assessment provide a stronger assessment when they are together than if used apart. Crossley (2012) summarizes it in this way, "In essence, scraping up the myriad evidential minutiae of the subcomponents of the task does not give as good a picture as standing back and considering the whole" (Crossley & Jolly, 2012, p. 33). Using a global assessment item, in addition to a list of performance items, show a likely

additive assessment effect, and removing too many of the items may have an unintended deleterious result.

Interestingly, similar to the mental cuing effect the items have on the global assessment, they also seem to be influencing the short answer questions. When asked to list one thing specifically done well, and one done poorly, raters often borrowed directly from the content of the 13 preceding items (see Tables 28 and 29). Content analysis showed even the wording of the items was preserved in many cases. There was some evidence from the feedback sessions that these short answer question mental cues also guide the rater when they communicate verbal formative feedback. For example, a surgeon commented that before having the OCAT in front of him, he focused solely on medical expert feedback. His feedback with OCAT stimulated reflection on documentation and communication skills as well. Therefore despite high inter-item correlations, and a small contribution of the items to the variance seen with ANOVA (6 percent only), the combination of the individual items and global assessment question can contribute to a merged assessment of performance and help generate targeted constructive feedback to specific elements of the task.

As an aside, staff raters received the global assessment item very well, finding it very helpful. In the feedback sessions they praised its ability to make them ‘take a side’ and believe it initiated more constructive criticism than would otherwise have been given, since they had to explain to the resident why they did not think they were ready to be independent (assuming they scored a ‘no’). Program directors also commented that the global assessment component of the OCAT would be helpful in a CBME program as it is of great concern if residents preparing to sit their RCPSC exams are not achieving ‘yes’ on the majority of forms. Raters who were working with more junior residents did feel conflicted about giving an answer of ‘no’ repetitively as it seemed harsh criticism. Therefore future rater training initiatives should continue to educate users on the appropriateness of junior residents receiving a ‘no’ until they are competent and independent.

Generalizability analysis. A generalizability study supported the conclusion that the OCAT is identifying differences in performance of individual residents by analyzing sources of variation (Table 17 and 18). The items on the form contributed minimally to the overall variance (6%), which is supported by their high inter-item correlations as was discussed above. A large contribution of the variance in the model came from the participants within their disciplines (34%). For example, while looking at residents just within orthopedics there were some high performers and some low performers as would be expected if the assessment tool were functioning correctly. The second major contributor of variance in the model was variance within the resident OCAT forms (33%). If an individual resident had three forms filled out, not all of the forms had the same scores (there were some high scores and some low scores given) likely reflecting the contextual nature of workplace-based assessment (Govaerts & Van der Vleuten, 2013) or differences in the raters. However, it is also possible that this is reflective of true differences in performance of the resident in context (Fisher, 2008; Strurman, Cashen, & Cheramie, 2005).

Assuming, however, that the variance seen within individuals is actually due to true differences between residents, on one hand it is good to see that staff surgeons are feeling free to rate the resident without a central tendency bias or only according to their PGY year. On the other hand, difference in scores for individual residents within a constrained time frame is a good reminder of why WBA tools like the OCAT should not be used as a stand-alone picture of competence. Only by aggregating multiple OCAT forms will the possible affects of social, contextual, and cognitive biases be mitigated (Schuwirth & Van der Vleuten, 2011; Van der Vleuten, et al. 2012). The data from the generalizability analysis indicate that to have a reliability coefficient of 0.74 (moderate), at least three forms would have to be completed for an individual resident. The submission of four forms increases the reliability coefficient to 0.79 and five submitted it becomes 0.83. From a program level perspective, five forms is a completely feasible number to collect over a two-month rotation block, especially if the resident is working with multiple staff in a variety of clinic settings (e.g. cancer clinic and benign disease clinic).

Relations with Other Variables Evidence

When considering how to gather evidence of OCAT performance relative to an already existing standard, there were not many options to compare to. No surgical clinics at the Ottawa Hospital were assessing their residents using a workplace based assessment tool. The majority of programs used the ITER to give the resident a retrospective end-of-rotation assessment in

direct CanMEDS language, with no specific mention of daily clinic performance. It would have been challenging to correlate the results of a new tool that was so intrinsically different from an ITER in structure, not to mention their relatively poor reliability performance (Regehr, Eva, Ginsburg, Halwani, & Sidhu, 2011; Ginsburg, Eva, & Regehr, 2013). Therefore, progression of performance scores on the OCAT was correlated with advancing year of training (PGY year) as a proxy for a gold standard tool.

There was a significant effect of PGY level seen with mean total item score differences between all levels except PGY4 and 5. The lack of difference between PGY4 and 5 is an interesting finding, potentially indicating that clinic competence for a generalist in a surgical specialty can occur before the final year of training. Also, breaking down the mean scores by PGY level is interesting in light of the entrustability behavioral anchors they correspond to. PGY1 and 2 residents (grouped because there were so few of them) had a mean score of 3.71 out of 6.0 (see Table 24 and 25). This corresponds to an average narrative rating between 3-“I had to help them from time to time” to 4-“I needed to be available just in case”. Considering that these are average ratings across items, and some items would have received scores pointing to constant guidance being necessary, the appropriateness of this mean score for junior residents is reasonable.

PGY3 residents received a mean score of 4-“I need to be available just in case”, and PGY4s and 5s received between 4-“Just in case” and 5-“I did not need to be there”. Although the difference in mean scores was less than a rating point of 1 (PGY3s 4.08 and PGY4/5s 4.92, for a difference of 0.84), this average point change illustrates an educationally relevant difference for a CBME assessment program. The entrustment decision is seen here and therefore it seems appropriate this would be applied to senior residents (PGY4 and PGY5s) on average more frequently than PGY3s. These differences between PGY years need to be studied in a larger population, especially to look at whether the OCAT can distinguish between PGY1 and 2 residents.

Few junior resident participants. A major limitation of the pilot study does pertain to the number of junior (PGY1 and 2) residents who participated in the study. The same number of junior residents were recruited for the study as senior residents (PGY4 & 5s), but fewer OCAT forms were submitted by these residents. There are multiple explanations for this finding that likely compound each other. The first, and most likely, is that junior residents attend clinic less frequently than senior residents, since they are often on “off-service” rotations (the OCAT was not used to assess them while “off-service”) and even while “on-service” they are more often busy with ward and emergency room duties. “Off-service” refers to an orthopedic resident doing a rotation in a surgical division that is not orthopedics, for example vascular surgery. Junior residents also switch rotations more frequently (one month rotation blocks rather than the three month blocks common for senior residents) and often have more rotations without many clinics (e.g. acute care surgery in general surgery).

Second, the OCAT pilot study was resident-driven, relying on residents to ask their staff surgeon to fill out a form at the end of clinic. It is possible junior residents are more intimidated by the staff surgeons and therefore nervous to ask them to focus on areas of feedback necessary. Or perhaps juniors do not have the same social relationship with the staff that senior residents enjoy. This could make them feel less comfortable interrupting staff at the end of a clinic to ask for an assessment. This question was not explored in the feedback sessions, but would be an interesting topic for future study.

Response Process Evidence

A large part of collecting response process evidence pertains to understanding what assessment tool users are thinking when they see and use an assessment tool. By considering their thought process, it is possible to control for some of the error resulting from test writers intending something that did not come across accurately, or inversely, test takers interpreting the meaning incorrectly. As well, it may be possible to use some of this evidence to guide future iterations of the tool to improve quality and reliability measurements. Response process also gathers evidence of the integrity of other aspects of the data collection process and feasibility concerns that may affect uptake of the tool (Downing, 2003).

Data collection accuracy. Steps were taken throughout the tool development process to ensure the data were both accurate and clean. The primary researcher and research assistant (C.S.) kept notes during the consensus groups and discussed the results with a third study team (N.D.) member after to ensure the list of final items was accurate. All pilot study data from returned OCAT forms was inputted in an excel file by one team member (C.S.) and double checked by the primary researcher against the original forms. The data was cleaned to remove OCAT forms with missing data and adjust any half values assigned. Future studies will include instructions in the rater training resources to only assign whole number values so staff are aware of the difficulty this causes during analysis. Fortunately during the pilot study half values were only assigned for a total of 10 items and so the arbitrary decision to round up was unlikely to have affected the results.

Effect of discipline. Of the three divisions involved in the pilot study, orthopedic surgery has had the most exposure to WBA tools using entrustability-aligned scales (Gofton et al., 2012), general surgery has recently used a similar tool, and obs/gyn has had no previous exposure. There was interest in looking at the results from the angle of difference between disciplines, hypothesizing that raters would be thinking about the scale differently due to different training exposures.

The g-study showed that discipline did not by itself account for a large amount of variability in scores between OCAT assessments, with mean scale totals showing a non-significant main effect size of 2.62 ($p = 0.08$). However, since this difference approached significance, post-hoc testing was completed and showed a significant difference in mean scores between obs/gyn and orthopedic surgery. Obs/gyn had a mean scale total score of 4.86 (S.D. 0.42) and orthopedics had 4.41 (S.D. 0.98) (see Tables 24 and 25).

From a response process perspective, it would be helpful to understand what raters from different specialties were thinking when they used the rating scale. Potentially orthopedic surgery raters are more accustomed to using the entrustability scale to assign lower scores as they have had more exposure to it (Gofton et al., 2011). However, an alternate explanation could be that obs/gyn residents actually achieve independent competence in their clinics at a faster pace given a narrower spectrum of pathology seen. Unfortunately the feedback sessions were only divided into residents and staff surgeons, with no record of which specialty participants originated from, so further comment on this issue cannot be made based on feedback group data at this time.

Effect of clinic type. The OCAT study stimulated discussion around the theoretical ideal academic surgeons should be pushing their residents towards. Many academic surgeons are sub-specialized and their clinics see patient complexity beyond what a graduated non-specialist surgeon in their field would be expected to manage. Instructions at the top of the OCAT told raters to rate their residents based on what a generalist surgeon in their specialty would be expected to know or do. At times the study team illustrated this concept by suggesting the rater should imagine him or herself testing this resident during a RCPSC final exam. Based on the resident's level of performance that day, regardless of the level of specialization of the clinic, the rater should ask himself or herself if that resident would pass the exam. In some cases, raters did not attend short rater training sessions held with the consent meetings, and the instructions at the top of the OCAT would be their only form of training received. If raters did not read the form properly, and used the entrustability scale incorrectly, it would be a major response process issue.

For this reason, the primary researcher was interested to know if there was an effect of which clinic the residents were being assessed in (specialist versus generalist clinic). The mean values differed significantly between clinics for items 1-11 with a mean of 4.41 (SD 0.86) given in generalist clinics and a mean of 4.67 (SD 0.64) given in specialist clinics, producing a significant small main effect (effect size 0.031) as seen in Table 26 and 27. Residents in specialist clinics were rated higher than those in generalist clinic on average. From the resident feedback session data, the opposite was expected. Prior to data collection, residents were concerned that staff surgeons would not read the instructions at the top of the form carefully and therefore rate them against their performance for that specialist clinic. If that were the case, one would think scores in specialist clinics would be lower since those patient cases are more difficult.

Previous studies with the mini-CEX and CAMEO have actually shown a similar effect with staff raters compensating for the difficulty of the cases by grading more leniently for the complicated cases (Norcini et al. 2003; Wilson et al., 2003). It may appear, based on this that case specificity bias is causing two residents with the same abilities to have differing scores, but it may also be argued that those with more advanced skills are entrusted with the more difficult cases. The numbers from this OCAT subgroup analysis were quite low, and therefore the true

relationship between ratings and clinic difficulty will have to be investigated further in future studies.

Six-point rating scale. WBA typically uses scales that are norm referenced (e.g. below average, average, and above average or unsatisfactory, satisfactory, and superior) as seen in a paper describing the ITERs used by the University of Toronto Faculty of Medicine (Ginsberg et al., 2013). In the OCAT project, an attempt was made to modify this usual way of writing anchor scales. The decision was made early in the project timeline to modify the 5-point O-SCORE operating room rating scale (Gofton et al., 2012) previously used at the study hospital, into a 6-point scale useable in clinic. It was thought that a six-point scale would: provide the opportunity to assess fellows in clinic, give the ability to rate residents as higher than competent for independent practice (i.e. recognize excellence), and push raters away from thinking of ratings of one to five as equivalent to the PGY levels one to five.

Looking at both quantitative and qualitative results, the six-point scale did not perform as well as intended. It was not good for recognizing fellows, as there was only one level (a '6') that essentially gave them a yes or no, without room for progression. As well, it is likely that some raters not assessing fellows interpreted six as 'independent for practice', which would have shifted the scores to the right. Looking at the mean score and standard deviation data from the O-SCORE, which are approximately one point lower than the OCAT (Gofton et al., 2012), it seems possible that this shift is occurring. In addition, the OCAT data does show that there are ranges for each item, implying there is no need to make supervisors rate away from PGY level because they are already doing it. Finally, some staff found it difficult to interpret '6' fellow level, whereas the '5' definition is well understood and actually the most crucial to know for residency training assessment in a CBME system. Therefore, based on pilot study results, it would seem prudent to consider changing the 6-point scale used in the OCAT to a 5-point scale for any future studies.

Feasibility. Time constraints in surgical clinic require that any additional paperwork introduced be short and efficient. The OCAT was noted to be a ‘time-added task’ but not ‘arduous’ in the feedback groups. Residents were positive about the response they received from staff when an OCAT was requested (it was easy to get them to fill them out). However the residents also suggested that they hesitated to ask staff to complete an OCAT in every clinic if the clinic had run over time, or the staff surgeon seemed to be in a hurry to go somewhere after clinic. Although the OCAT took only five to ten minutes to complete and surgeons thought it was not onerous, they did agree that if possible the legal sized sheet should be shortened to a regular sized sheet.

Staff surgeons also said the items were excellently worded and very applicable to the way they think about competence in surgical clinic. They did not think any major items were missing that they wanted to assess. When asked if they would like to change anything about the OCAT, they volunteered only the wording of the time management item and nothing else.

There was concern, however, about what would happen to OCAT feasibility outside of study conditions when residents no longer got paid to remind staff surgeons to fill out a form (the ten dollar coffee voucher they received). A new mechanism for OCAT distribution and collection will have to be created moving forward now that the pilot study is completed if the OCAT is going to become a regular WBA tool.

A few of the staff surgeons were reluctantly acceptant of the assessment tool since they acknowledged that increased workplace based assessment will be what is required in a CBME program. An example of this hesitance is illustrated in the following quote from a program director participating in the feedback group,

The reality is we’re going to see more and more of these forms [assessment tools] because the college [RCPSC] will require them and we’ll be having to figure them out. I think that this is actually quite a good form compared with others... It is time-consuming, but it isn’t arduous. [Program Director]

This is likely to be a common attitude as the formative assessment culture slowly shifts over the next few years in surgery.

Tool characteristics impacting the response process. It may be apparent by this point that these response process issues are largely due to the involvement of subjective raters. Clinical supervisors are an obligatory part of WBA, but as previously discussed (*Biases in surgical workplace-based assessment*) the judgment of these raters is prone to bias (Williams et al., 2003, Yeates et al., 2013). Interestingly, recent work suggests that raters do not so much disagree on trainee performance, but rather on how to interpret the assessment scale and how to “choose a number” (the response format) (Crossley et al., 2011). How to improve rater agreement in WBA and meaningfully capture subjective judgment is currently not clear. Within this project, some steps were taken to improve rater response process including: using a global assessment item to capture the subjective impression of the rater (*discussed under Internal Structure evidence*), incorporating a construct-aligned narrative anchor scale based on entrustability, and considering how best to train the raters to use the tool.

Entrustability-aligned rating scales. This study used a 6-point rating scale with construct-aligned narrative entrustability anchors. As previously discussed, assessment tools that include entrustability decision-making show promise for improving documentation in CBME. The goal of entrustability anchors is to align an assessment tool with a clinical rater’s internal thought process.

There has been some discussion recently to indicate that entrustability anchor scales may help bridge a gap between the assessment judgments of clinical supervisors and their WBA instruments. Entrustment is a necessary process through which a learner must navigate on the road to professional competence (Mann, 2011) and clinical supervisors are accountable for helping residents shoulder increasing responsibility. These supervisors should continually be asking themselves if the resident is capable of completing a task independently (Mulder, ten Cate Daalder, & Berkvens, 2010; Ten Cate, 2005).

Rather than asking a staff physician to translate these complex assessment thoughts into ordinal categories such as “above average, below average, etc.”, entrustability scales acknowledge the categorical judgment – to trust or not to trust – raters have been shown to use in the workplace (Yeates et al., 2013). Since some rating scale error is due to failure to commit the implicit categorical (interval) judgments of workplace evaluation into the ordinal judgments traditionally required by abstract scales (Macrae & Bodenhausen, 2000), reverse engineering descriptors to fit raters’ existing categorical schemas can increase the reliability of assessments (Crossley & Jolly, 2012). For example, when asked to judge whether a trainee required direct,

indirect, or distant supervision with a case, anesthesia staff physicians were shown to be much more reliable in their assessments than when they were asked to judge what is expected of a trainee at different stages of training (only 9 assessments required versus 50 to reach significance) (Weller, Jones, Merry, Jolly, & Saunders, 2009). Evidence from this OCAT pilot study, and the O-SCORE (Gofton et al., 2012), demonstrate a similar effect with only five OCAT assessments needed to generate a reliability of 0.83.

Narrative comments from staff physicians using the OCAT support the intuitive nature of the entrustability scale and its ease of use. Further study will have to be completed however before the entrustability scale's effect is completely understood. Currently there is minimal to no evidence of how residents perceive the entrustability scale, compared to a more traditional norm-referenced one. As well, there is also a question of how entrustability scales affect the need for the type of rater training given when a tool is introduced.

The place of rater training. Whenever a new WBA tool is implemented, concerns over rater bias and subjectivity should prompt discussions of how much rater training is necessary (Williams, 2003; Dath & Iobst, 2010). There are mixed conclusions in the literature regarding the benefit of rater training. Some groups have shown that rater training can increase proper use of an assessment tool (Angkaw, Tran, & Haaga, 2006; Woehr & Huffcutt, 1994). However these training sessions are resource intensive, time consuming (often requiring two to four hours of a rater's time), and are difficult to implement outside of a study environment (Holmboe et al., 2011). In contrast, a randomized controlled trial evaluating the effect of a training workshop (compared to no intervention) on inter-rater reliability and accuracy of ratings found no difference between scoring in the groups (Cook, Dupras, Beckman, Thomas, & Pankratz, 2008).

Whether or not rater training is helpful to decrease rater bias, the length and detail required by many literature training success stories just are likely to fail in the long term in a surgical environment. Time to continually train staff on existing and new tools is in short supply, and for this reason staff attendance at rater training sessions is notoriously poor (Dudek et al., 2012). Overall, it is more reasonable to design assessment tools that are practical and easy to understand, than use complex rater training. Crossley's work on construct-aligned scales, and the success of other entrustability-aligned anchor scales in helping raters understand what they are being asked to rate (Gofton, 2012; Warm et al., 2014; Weller et al., 2014), suggest that traditional full-length rater training programs may not be necessary for the OCAT.

That said, when using entrustability scales for the first time, education leaders will still

need to monitor for those raters failing to use the ‘external reference’ as intended (i.e. independent competence, not peer group comparison). There was some evidence of a failure to use the external reference as noted by the residents in the OCAT feedback groups (see resident quotes on p.64 of *Results*). A method of creating targeted rater re-orientation, called frame-of-reference training, could be attempted by monitoring regularly for outlying raters likely to benefit from re-alignment to scale principles. The application of one such method, called frame of reference training was recently to train raters for procedure skills OSCE (Pugh et al., 2015).

The OCAT used two forms of rater training that were based on feasibility, not on the gold standard for rater training in the literature (Holmboe et al., 2011). The first was brief rater orientation at consent meetings and second, the written instructions on the top of the tool. Within the instructions on the top of the tool, users received two key pieces of information: that they would be scoring using an entrustability scale and that they should score from the point of view of a generalist in their specialty. Some evidence from this study points to a need for more extensive rater training, including conflicting opinions of residents on whether the staff surgeons were completing the entrustability scale from traditional norm-referenced perspective and questions from staff raters during the pilot study on how to use the tool in a specialty clinic. Further study is needed to determine the extent of rater orientation or formal training that is necessary for proper completion of the OCAT and its entrustability rating scale.

Consequential Evidence

Does the OCAT change formative feedback given in surgical clinic? Quality feedback from assessors appears to be at the heart of the assessment process in WBA (Bok, 2013). A meaningful piece of consequential evidence that became apparent throughout this pilot project was the impact the OCAT had on the occurrence of daily formative feedback. At consent meetings prior to start of the pilot, residents were asked to respond to a question enquiring whether they receive appropriate feedback on their performance in surgical clinic. Fifty-two of the residents responded to the survey (65%) and the majority of the residents were either ‘neutral’ towards the comment, or disagreed (40/52) (see Table 12). Only 12 of 52 residents agreed or strongly agreed with the statement and seven of these were from one specialty (obstetrics and gynecology). Written comments at the end of the survey (Table 13) line up with this observation and are indicative of the rare occurrence of formative feedback and the need for a formative feedback tool like the OCAT.

If there is evidence to say that feedback occurs infrequently, there is even more evidence to support a lack of constructive negative feedback (Dudek et al., 2005; Hanson et al., 2013). In fact, evidence exists to suggest that final assessments are often not consistent with a rater's actual judgment, especially for poor performers (Dudek & Dojeiji, 2014). Rating forms, including sections with written comments, are often not used at all or are filled with broad statements such as "good job" or "keep reading" (Canavan, Holtman, Richmond, & Katsufakis, 2010). A similar phenomenon was obvious in the short answer responses on the OCAT. Question 14 asked raters to give one *specific* thing done well in clinic (only one response was asked for to facilitate easy reply). This item was completed on 92% of the forms but analysis showed the majority of these were very general, vague comments. Only five comments were coded as *specific*, despite explicit request for this (see Table 28 for specific comments).

Question 15, a request for one *specific* suggestion for improvement did not have as high a response. Only 96 of 132 comments were complete, with 27% left completely blank. Including "none" or "nothing" statements (13%), veiled positive comments (e.g. "work on differential, but actually very good for level") (17%), and "read more" (8%), 65% of the time no constructive criticism was actually given. Finally, to compound the issue, 31 forms (36% of comments made) contained very general comments (e.g. keep increasing knowledge base, keep at it, etc.). Obviously, raters struggled to deliver creative, constructive narrative critiques of their residents. To be fair, however, there were 49 responses given that could be correlated directly to items proceeding on the OCAT as previously discussed, with some forms containing more than one item to improve on. See Table 29 for a breakdown of which items were most frequently used.

There has been a recent push in assessment to consider the necessity of narrative comments for WBA (Govaerts, van der Vleuten, Schuwirth, & Muijtjeijns, 2005; Pelgrim, Kramer, Mookink, & van der Vleuten, 2012). But if raters are not taking the opportunities presented to them to do this, how can we move forward? Bok et al (2013) talk about how the format of assessment tools themselves may be giving raters an excuse for refraining from narrative qualitative feedback, by requiring a quantitative mark first. By the time surgeons reach the end of the OCAT, they may feel they have already done their duty by completing the previous 13 items plus the global assessment item and not realize the necessity of the short answer questions. However another likely explanation for the paucity of constructive critique seen on the OCAT is one of its strongest consequences, illuminating how few academic staff

surgeons know how to convey areas needing improvement to their residents. Giving good feedback is a skill that can be developed with training (Brown & Cooke, 2009).

Teaching staff raters to convey insightful feedback to their residents should be a priority in professional development. Most clinical teachers are not properly trained to give quality feedback, so it is important to stress the following definition, 'Feedback is *specific* information about an *observed* performance (compared to a standard), given with the *intent to improve* performance' (Van de Ridder, Stokking, McGaghie, & ten Cate, 2008). Consistent and systematic feedback, delivered by a credible and authoritative source, can positively impact clinical performance and may be one of the most powerful influences helping learners progress (Veloski, Boex, Grasberger, Evans, & Wolfson, 2006).

Feedback has been shown to be beneficial in a variety of domains. There is evidence for its improvement of diagnostic accuracy (Wigton, Patil, & Hoellerich, 1986), documentation (Opila, 1997), professionalism ratings (Brinkman et al., 2007), resource allocation (Studnicki, Bradham, Marshburn, Foulis, & Straumfjord, 1993), and even communication skills (Brinkman et al., 2007), for which it has also been consistently linked with improved health outcomes. A meta-analysis by Hattie & Timperley (2007) involving 196 studies showed an average effect size of feedback to be 0.79, twice that of the average effect size of 'education'. Obviously strengthening formative feedback by introducing feasible tools into the workplace can greatly affect the strength of a residency assessment program.

The OCAT definitely improved formative feedback given and received. Residents and staff unanimously agreed in the feedback session that the biggest impact of having the OCAT form available was that it was a reminder to give feedback. Not only was a window of time created for feedback, the quality of the feedback also improved. One resident commented that the feedback was specifically more constructive when it was guided by the elements of clinic competence indicated in the items. The OCAT also generated a more holistic discussion of clinic performance than one strictly focused on more traditional medical knowledge components of performance. From a CBME perspective this is helpful, since the inclusion of multiple CanMEDS competencies in feedback is essential for this system.

Finally, one of the key functions of the OCAT compared to the currently used ITER is its ability to give feedback in real time (Wagner et al, 2014). Change in behavior is promoted based on recent, memorable examples at the time of the assessment, as noted by this resident when describing the benefits of the OCAT,

Also, it [the OCAT assessment] was real time. Rather than being three weeks down the line when I'm done and can't respond to it. If they give me something to work on, then I can show them I'm trying to improve it.

Formative feedback and trust. This resident seems to be recognizing that assessment and feedback are two-way streets, both the responsibility of the assessor and also the person being assessed. There can be significant tensions in the clinical environment between maintaining collegiality and ease of communication between the staff giving the feedback and the resident receiving it. This, and the resident's willingness to seek feedback for the same reasons (Eva et al., 2012), can prevent honest communication from occurring (Kogan et al., 2012; Mann, 2011). In addition, there is evidence that it is easier for residents to accept constructive feedback and know what to do with it when they receive it in the context of a mentorship relationship (Telio, Ajjawi, & Regehr, 2014).

From the perspective of a social constructivist, if there is no trust in the assessment relationship, students are less likely to receive the criticism as accurate and learn from it (Weller et al., 2014). "Trust allows the trainee to experience increasing levels of participation and responsibility in the workplace in a way that builds competence for future practice," say Hauer et al. (2014, pg.435). Cruess & Cruess (2014) talk about two types of entrustment decisions, ad-hoc and summative. Ad hoc entrustment decisions are built on presumptive trust (prior credentials such as an MD) and initial trust (essentially first impressions). The trust that contributes in a helpful way to competency progression however, needs to be built on more. Cruess & Cruess term this "grounded trust", based on systematic data collection over time by multiple assessors. The OCAT has the potential to contribute to grounded trust between surgical residents and their staff supervisors, especially since they spend time together intermittently over five years. If improved formative feedback is given over this time in the context of a strong mentorship alliance, residents have the potential as adult learners to maturely accept constructive feedback and move confidently towards competence.

How do residents feel about receiving low scores on the OCAT? As adult learners residents are professional students who, to get where they are, excel at writing tests and scoring well. Another reason it is difficult for residents to accept formative assessment feedback (besides a lack of trust) is that they may be preconditioned by a culture of high-stakes assessment (Jones et al., 2014). Having to achieve the best grades in the class to get into medical school can

cause difficulty down the line when they receive lower scores they are unaccustomed to. Understanding this, staff physicians may feel the need to preserve a positive working-relationship and skew assessment results to the top of the rating scale (Bennett, 2011). On the other hand, it is important not to under-sell a residents' ability to desire constructive feedback as evidenced by this quote from a PGY4 resident in the feedback group,

I always think that it's better when someone actually points out a way to be better. I find the ones where someone says, "Oh, you're a five". And I'm like, "really, I know that's not the case". I know they don't feel that way and I think, "Why don't you just write down how you really feel and we can make this useful".

On a practical level, there are many reasons (including risk of litigation and more paperwork) why clinical raters are unlikely to tell their residents that their performance was "unsatisfactory" and are even more unlikely to fail them (Dudek et al., 2005; Weller et al., 2014). However, it is possible that an entrustability scale can make lower scores more acceptable. Rating scale assessment values of 1 and 2 would traditionally denote "below average" or really, failing performance. However on an entrustment scale they simply convey what happened from the perspective of the rater; "I had to do", or "I had to walk them through it".

In effect, entrustability scales help staff raters make non-pejorative assessments based on narrative descriptors that reflect real-world judgments. This increases the likelihood of an honest assessment (i.e. it allows for a more valid assessment of the construct of interest) and decreases the effect of leniency bias and the 'Mum effect' (Yeates et al., 2013). It also helps the trainee interpret the assessment as a representation of their progress towards safe independent practice, rather than as a comparison with their peers or with an abstract construct such as their PGY level. Work on the O-SCORE (Gofton et al., 2012) found that residents were comfortable receiving lower scores when the scale they were being compared against was worded in terms of entrustability. The feedback session results from the OCAT formed the same conclusion. Five out of six residents agreed that lower scores were more acceptable on the OCAT rating scale. As this PGY1 said, "I don't find a one or two offensive because the scale is not saying I'm totally incompetent, only that I needed help".

Extrapolating from these results, it seems plausible that a component of the success of entrustability scales is that they focus trainee attention on the practical goal of independent practice, rather than the personal or psychological goals of approval and competitive excellence. This is a point to study in the future when delving deeper into how trainees think about

entrustability. In light of CBME's reliance on formative feedback to guide residents between milestones (Holmboe et al., 2010), creating this culture of acceptance for low stakes daily assessments is essential (Schuwirth & Ash, 2013; Newton, Henderson, Jolly, & Greaves, 2015).

Residents in difficulty and the OCAT. A final piece of consequential evidence stemming from the pilot project is considering how the OCAT can help identify residents who are having difficulty progressing towards clinic competence. In practical terms this will mean helping program directors determine whether to hold a resident back or not, and whether remediation plans are necessary. One of the visions for CBME is early identification of residents in difficulty. Most residents desire this early intervention, as demonstrated by this quote from an orthopedic PGY3, "It's not good [receiving a one or two on the rating scale]. But if it's honest feedback, and you have time for constructive conversation about how to improve, then it's valuable."

Promising research in social work trainees has shown that assessment forms using narrative descriptions of performance (similar to entrustability anchor descriptions) may be better at identifying borderline performance than traditional forms (Regehr, Bogo, Regehr, & Power, 2007). Also, entrustment scales can provide a way to track the learner over time, and show where they 'fall off the curve' compared to expected level of clinical entrustment (Weller et al., 2014). Residents in the feedback session re-iterated the OCAT's ability to point out these areas of deficiency and most appreciated the timeliness of the assessment. Also, having the context of that day's clinic helps because it allows for examples to be given. Residents buy in more because there are 'discrete examples' of areas of strength or weakness.

The dichotomous global assessment item at the end of the OCAT can also be used to derive an 'independence curve', documenting a transition from 'No' to 'Yes' for each resident across the years of training. Even for residents showing normal progression on these long-term performance curves, there will be variability in the entrustability growth curve over time, reflecting contextual variations. However, compared to cohort measurement, which would measure only changes in standing, this achievement model unambiguously reflects individual learning.

Limitations of Study

Context limitations. A significant limitation the OCAT shares with all WBA tools is its inability to completely account for context complexity (Weller et al., 2014). Despite assessments being individualized to a particular resident, participating in a hepatobiliary clinic is still a different task to complete than a follow-up hernia clinic (but encompassed by the same EPA). This has been referred to and discussed in the literature as case specificity (Elstein, Shulman, & Sprafka, 1981) or context specificity, meaning that success in one context is specific to that context, and cannot be assumed to apply to the next (Norman, Bordage, Page, & Keane, 2006).

There are many contextual factors influencing entrustment decisions (Hauer et al., 2014), and therefore multiple assessment forms, ideally completed by multiple raters, need to be filled out to see this effect dissipate. Formative WBA should be viewed as useful in the moment to guide residents in their areas of strong and weak performance, but only useful to the residency program in the long-term, once trends can be determined over time. Isolated use of the OCAT would involve fluctuations due to uncontrollable external contextual factors.

Also, if truly taking the social elements of context into account, WBA in surgery should involve other members of the surgical team (Hodges, 2013). Resident function in the workplace context can no longer be seen as an individual activity. Clinical care in hospital and in outpatient settings is usually a team responsibility. Therefore, it is difficult to isolate and appraise the contributions and clinical competence of individual practitioners (Noel et al., 1992). Lorelei Lingard refers to the importance of considering ‘collective competence’, rather than just individual competence (Lingard, Espin, Evans, & Hawryluck, 2004). The team may struggle due to an individually incompetent member, and on the other hand, weakness in trainee knowledge and skill may be masked by the contributions of others (Williams et al., 2003).

Assessment from multiple perspectives would also add strength and a holistic perspective to surgical clinic WBA (Hodges, 2013). The OCAT only assesses resident performance from the expert surgeon supervisor’s viewpoint. This is important, however richer information can potentially be gathered by actively involving interdisciplinary staff in the assessment process. Strong data exist for the validity of scores resulting from multisource feedback tools (Donnon, Al Ansari, Al Alawi, Violato, 2014) and they have become a valuable assessment method. Multisource feedback tools are logistically difficult to implement, but once they are part of the assessment culture have the ability to add an interesting peripheral vision to that of staff surgeons in clinic (Epstein, 2007).

Additional limitations. Other limitations of the OCAT pilot project have the potential to be improved in future work. First, as previously mentioned, there were only 19 junior residents who submitted OCAT forms during the pilot project limiting the conclusions that can be drawn for this population. More work will need to be done to explore why they did not have forms filled out, and to ensure that the entrustability rating scale is a useful measure to help them progress from junior to senior levels. During the next iteration of the OCAT study, intentional targeting of junior residents should occur to ensure an adequate sample size.

Second, a minor limitation of the consensus group process was the lack of an anonymous method of voting (e.g. holding up different colors of cards). Due to the study team's goal of stimulating discussion to understand thought process, a fuller number of opinions may have been sacrificed. Anecdotally it seemed as if all group members were able to voice their opinion, but obviously it is not possible to know completely.

Third, due to feasibility constraints, it was impossible to blind raters to the PGY level of residents they were assessing. Even in the largest programs, staff surgeons usually know how long an individual resident has been around the program. Given that the OCATs relation to other variables evidence is based on progression seen through PGY levels, it would be more convincing if blinded external staff experts could be brought in to work with the resident in clinic for a day and then give an assessment. However, in the real clinical environment, surgeons who know their residents well use WBA tools. Even if the tool was shown to have validity evidence by blinded assessors, one could never know if familiar surgeons are using it in the same way as the blinded surgeons did during the study. At this time, the generalizability analysis indicates staff surgeons are not scoring according to PGY level since the variance in individual resident's scores is high (Friedman factor in Table 17). However in the future, once more WBA tools are implemented as part of CBME, the OCAT could be evaluated against them to see how they compare.

Strengths of the OCAT

Despite these limitations, it is important that they not overshadow the strengths of the project. Significant methodological rigor was applied throughout the tool development process to gather multiple sources of validity evidence, setting the OCAT apart from many previously published tools in the literature.

It is fair to say with certainty that the content of the OCAT items is strong. All surgical divisions both locally and nationally, and both surgeon and resident opinion, were sought in

order to make sure no voices were missed. At all crucial decision-making points of the project, such as writing the descriptors for the items, two or more team members were involved. This team approach makes the OCAT a thoughtful piece of work, incorporating the assessment expertise of multiple authorities.

This thoughtfulness also extends to how feasible the OCAT is. It can be immediately implemented in a surgical clinic environment since it was developed for the reality of clinic now, not a theoretical ideal under a fully developed CBME program. From the eye-catching green color of the form, to the limited time it takes to complete, the pilot project demonstrated the OCAT will be utilized. Even at the end of a busy surgical clinic, it is possible for a surgical supervisor to sit down and provide quality formative feedback that was previously missing.

Future Directions

This pilot project has collected both psychometric evidence and qualitative commentary indicating that the OCAT provides ratings that are valid and that it is a feasible method of assessing surgical resident performance in clinic. However, the findings of this study reveal a plethora of validity issues for further research, including but not limited to: how to aggregate multiple OCATs, how much rater training is necessary, whether the OCAT should be paper-based or electronic, and how many items are optimal to have on the tool.

Future work would see the OCAT streamlined for feasibility based on pilot data. Based on quantitative and qualitative feedback, changes that could be made include: removing items 1 and 11 (clinic preparation and attitude), adding a dichotomous professionalism question to replace ‘attitude’, changing the rating scale to a five-point scale, modifying the orientation of the yes/no response option to the global assessment item (based on work by Streiner & Norman, 2008), and modifying the wording of the instructions to make it clear that residents should be rated as generalists in their specialty. Further validity evidence in a wider sample of surgical residency training programs could then be gathered.

How are these OCATs going to be aggregated meaningfully? If WBAs contribute bits of information to the overall picture, human judgment is still crucial for coalescing assessment information from various sources (Bok et al., 2013). The current Canadian system requires Program Directors, in partnership with their program training committee, to combine assessments into an overall progression decision (and ultimately the FITER) at the end of residency training. This is unlikely to change and is why Hodges says, “the assessment of the future must be an aggregation” (Hodges, 2013, p.567). The g-study indicates there was acceptable reliability achieved when a minimum of five OCAT forms were collected during the pilot study. However, how varying results of these 5 or more OCATs are aggregated will have to be left up to human judgment. The OCAT should also be used in combination with other assessment tools and best practice still has to be defined.

What is clear is that when new tools are introduced into an assessment program, information on how the tool will contribute (or not contribute) to final progression decisions should be shared with staff and residents clearly before they use it. This will help ensure that formative assessments are not misperceived as summative. If students and supervisors interpret the value of individual low-stakes assessment in the same way, residents and raters both will be better able to focus on the potential learning value of WBAs, rather than on their summative consequences (Bok et al., 2013).

How much rater training? This OCAT pilot project found the tool to display sound validity evidence even with minimal rater training conducted. However, as noted above, some issues came up related to response process and especially the use of an entrustability scale for the first time in some divisions. A follow-up study is required to further investigate how much rater training in the traditional sense is necessary for use of the OCAT. Potentially just using the written instructions on the tool or a frame-of-reference training booklet may be possible as raters accommodate to this new WBA tool. It would be interesting to note participant comments and thoughts along the way regarding the entrustability scale and the subjective global assessment item.

Paper or Electronic OCAT? Transitioning assessment tools from paper-based forms to electronic documents that may be simpler to aggregate and track in the long run is not a simple process. Users want the electronic form to be as simple to use as a paper one, but electronic forms inevitably come with login passwords and extra steps (Dudek, Papp, & Gofton, in press). Concerns over loss of narrative data in short answer questions have also been voiced, since typing responses is often more difficult on a tablet than it would be if one was handwriting (Saleem et al., 2009).

When asked whether they would prefer the OCAT to be paper-based or electronic, the study team expected residents to be in favor of the electronic version and staff surgeons to be opposed to it. Residents more commonly have their iPads accessible and tend to be more familiar with online assessment forms (required from their University). Interestingly, the OCAT feedback sessions found that staff surgeons said they would prefer an electronic form and residents said they would prefer it stays on paper. Residents seemed to recognize that technological issues (e.g. failure to login or getting ‘buried’ in an inbox) often prevent the staff from completing forms. Whereas when the form is on paper, residents can hand it to them and then not leave until it is filled in.

The next iteration of the OCAT study should likely remain on paper to keep as many factors consistent while varying the amount of rater training. However, as CBME assessment programs become more complex and involve multiple daily WBAs, a solution of how to aggregate all the forms will be necessary. An electronic form would be much more helpful in this regard (Buzzetto-More & Alade, 2006). Finally, electronic platforms have even more ability to stimulate timely aggregated feedback because assessments can be inputted the same day and appear immediately for the residents’ review (Wagner et al., 2014). There have been a few attempts to create streamlined systems able to produce performance curves. A system like this could be used with the OCAT to compare residents’ performance to see where individuals are falling off the curve and may need extra remedial support.

Additional short and long-term suggestions for future study mentioned in this paper have been aggregated and can be found in Appendix S.

Conclusion

Workplace-based assessment provides the opportunity for surgeon supervisors to deliver direct, targeted feedback to their residents within the authentic social context. Following reception of this feedback, surgical residents have a responsibility as adult learners to use it to continue developing independent skills and behaviors in the clinical environment. Multiple sources of validity evidence have been collated here to support that with this preliminary information the Ottawa Clinic Assessment Tool (OCAT) is a valid, reliable, and feasible way to document this progression towards competence in surgical clinic. Further iterative validity evidence will need to be collected over time moving forward. Currently however, it seems that the OCAT fits into a competency based assessment system, requires less rater training than traditional tools and more ability to convey subjective opinion due to its entrustability anchor scale, and promotes a culture of acceptance for constructive criticism. Attentiveness to the likelihood for rater bias and contextual influences mean that individual OCATs should not be used in isolation, or as a stand-alone summative assessment. Instead, the OCAT should continually challenge surgeon raters to sample daily performance data and translate these judgments into productive, documented feedback for the residents in their community of practice.

Acknowledgements

The primary author would like to acknowledge the help of Christine Seabrook (C.S.), Dr. Nancy Dudek (N.D.), Dr. Stan Hamstra (S.H.), Dr. Wade Gofton (W.G.), Dr. Tim Wood and Dr. Eric Dionne. Ms. Seabrook, the surgical education Research Coordinator at the Ottawa Hospital, for her help with the consensus group process, data management/storage, and participation in analysis of the short answer question data. Dr. Dudek for her facilitation of the consensus groups as well as the feedback groups. Dr. Hamstra and Dr. Gofton for their ongoing discussion of project details including specifically the National expert survey. Finally Dr. Tim Wood and Dr. Eric Dionne for their psychometric expertise and comments on the manuscript.

References

- Alman, B, Ferguson, P, Kraemer, W, Nousianien, M, Reznick, R. (2013). Competency-based education: a new model for teaching orthopedics. *Instructional Course Lectures*, 62, 565–9.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Angkaw, A, Tran, G, Haaga, D. (2006). Effects of training intensity on observers' rating of anxiety, social skills, and alcohol-specific coping skills. *Behav Res Ther*, 44, 533–544.
- Arora, S, Sevdalis, N, Suliman, I, Athanasiou, T, Kneebone, R, Darzi, A. (2009). What makes a competent surgeon?: Experts' and trainees' perceptions of the roles of a surgeon. *American Journal of Surgery*, 198, 726–32.
- Baker, K. (2011). Determining Resident Clinical Performance: Getting Beyond the Noise. *Anesthesiology*, 115(4), 862–878.
- Bancroft, G. N., Basu, C. B., Leong, M., Mateo, C., Hollier, L. H., & Stal, S. (2008). Outcome-based residency education: teaching and evaluating the core competencies in plastic surgery. *Plastic and Reconstructive Surgery*, 121(6), 441e–448e. <http://doi.org/10.1097/PRS.0b013e318170a778>
- Bandiera, G, Lendrum, D. (2008). Daily encounter cards facilitate competency-based feedback while leniency bias persists. *CJEM*, 10(1), 44–50.
- Beckman, T. J., Cook, D. A., Mandrekar, J. N. (2005). What is the validity evidence for assessments in teaching? *Journal of General Internal Medicine*, 20, 1159–1164.
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. <http://doi.org/10.1080/0969594X.2010.513678>
- Bhatti, N, Cummings, C. (2007). Competency in surgical residency training: Defining and raising the bar. *Academic Medicine*, 82(6), 569–73.
- Bogo, M, Regehr, C, Woodford, M, Hughes, J, Power, R, Regehr, G. (2006). Beyond competencies: field instructors' descriptions of student performance. *Journal of Social Work Education*, 42(3), 579–593.
- Bok, H. G., Teunissen, P. W., & Favier, R. P. (2013). Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Medical Education*, 13(1), 123. <http://doi.org/10.1186/1472-6920-13-123>

- Brennan, B. G., Norman, G. R. (1997). Use of encounter cards for evaluation of residents in Obstetrics. *Academic Medicine*, 72(10), S33-34.
- Brinkman, W, Geraghty, S, Lanphear, B, Khoury, J, Gonzalez del Rey, J, Dewitt, T, Britto, M. (2007). Effect of multisource feedback on resident communication skills and professionalism: a randomized controlled trial. *Arch Pediatr Adolesc Med*, 161(1), 44-49.
- Brooks, M. a. (2009). Medical education and the tyranny of competency. *Perspectives in Biology and Medicine*, 52(1), 90-102. <http://doi.org/10.1353/pbm.0.0068>
- Brown, N., & Cooke, L. (2009). Giving effective feedback to psychiatric trainees. *Advances in Psychiatric Treatment*. 15(2), 123-128. <http://doi.org/10.1192/apt.bp.106.003293>
- Buckley, S, Coleman, J, Davison, I, Khan, K, Zamora, J, Malick, S, et al. (2009). The educational effects of portfolios on undergraduate student learning; a best evidence medical education (BEME) systematic review. *Medical Teacher*, 31, 282-98.
- Buzzetto-More, N, Alade, A. (2006). Best practices in e-assessment. *Journal of Information Technology Education*, 5, 251-69.
- Campbell, S. M., & Cantrill, J. a. (2001). Consensus methods in prescribing research. *Journal of Clinical Pharmacy and Therapeutics*, 26(1), 5-14. <http://doi.org/10.1046/j.1365-2710.2001.00331.x>
- Canavan, C, Holtman, M, Richmond, M, Katsufakis, P. (2010). The quality of written comments on professional behaviors in a developmental multisource feedback program. *Academic Medicine*, 85, S106-S109.
- Carraccio, C., & Burke, A. E. (2010). Beyond Competencies and Milestones: Adding Meaning Through Context. *Journal of Graduate Medical Education*, 2(3), 419-422. <http://doi.org/10.4300/JGME-D-10-00127.1>
- Chou, S, Cole, G, McLaughlin, K, Lockyer, J. (2008). CanMEDS evaluation in Canadian postgraduate training programmes: tools used and programme director satisfaction. *Medical Education*, 42(9), 879-886.
- Cleland, J. a., Knight, L. V., Rees, C. E., Tracey, S., & Bond, C. M. (2008). Is it me or is it them? Factors that influence the passing of underperforming students. *Medical Education*, 42(8), 800-809. <http://doi.org/10.1111/j.1365-2923.2008.03113.x>
- Cogbill, T. H., Malangoni, M. A., Valentine, R. J., & Potts, J. R. (2014). The General Surgery Milestones Project. *Journal of the American College of Surgeons*, 218(5), 1056-1062. <http://doi.org/10.1016/j.jamcollsurg.2014.02.016>

- Cook, D. a., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine*, 119(2), 166.e7–16.
<http://doi.org/10.1016/j.amjmed.2005.10.036>
- Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2008). Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *Journal of General Internal Medicine*, 24(1), 74–79. <http://doi.org/10.1007/s11606-008-0842-3>
- Cook, D, Beckman, T, Mandrekar, J, Pankratz, V. (2010). Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability. *Adv in Health Sci Educ*, 15(5), 633–645.
- Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90(2), 218-244.
- Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Medical Education*, 45(6), 560–569.
<http://doi.org/10.1111/j.1365-2923.2010.03913.x>
- Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. *Medical Education*, 46(1), 28–37. <http://doi.org/10.1111/j.1365-2923.2011.04166.x>
- Cruess, R, Cruess, S. (2014). Professional Trust. In *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society*. Hoboken, NJ: John Wiley & Sons.
- Cruess, R, McIlroy, J, Cruess, S, Ginsburg, S, Steinert, Y. (2006). The professionalism mini-evaluation exercise: a preliminary investigation. *Academic Medicine*, 81(10Supp), S74–8.
- Dall’Alba, G. (2009). Learning professional ways of being: Ambiguities of becoming. *Educational Philosophy and Theory*, 41(1), 34–45. <http://doi.org/10.1111/j.1469-5812.2008.00475.x>
- Darosa, D. A., Zwischenberger, J. B., Meyerson, S. L., George, B. C., Teitelbaum, E. N., Soper, N. J., & Fryer, J. P. (2012). A Theory-Based Model for Teaching and Assessing Residents in the Operating Room. *Journal of Surgical Education*, 70(1), 24–30.
- Dath, D., & Iobst, W. (2010). The importance of faculty development in the transition to competency-based medical education. *Medical Teacher*, (905), 683–686. <http://doi.org/10.3109/0142159X.2010.500710>
- Davies, H, Archer, J, Southgate, L, Norcini, J. (2009). Initial evaluation of the foundation assessment programme. *Medical Education*, 43, 73–81.

- Davis, D, Lee, G. (2011). The use of standardized patients in the plastic surgery residency curriculum: teaching core competencies with objective structured clinical examinations. *Plastic Reconstructive Surgery*, 128(1), 291–298.
- De Champlain. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, 44, 109-117.
- Donnon, T., Al Ansari, A., Al Alawi, S., Violato, C. (2014). The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Academic Medicine*, 89(3), 511-516.
- Downing, S. M. (2003, September). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37, 830-837. <http://doi.org/10.1046/j.1365-2923.2003.01594.x>
- Downing, S. M. (2004). Reliability : on the reproducibility of assessment data. *Medical Education*, 38, 1006–1012. <http://doi.org/10.1046/j.1365-2929.2004.01932.x>
- Downing, S. (2006). Twelve Steps for Effective Test Development. In T. Downing, Steven, Haladyna (Ed.), *Handbook of Test Development* (pp. 3–26). Routledge.
- Dreyfus, S.E., Dreyfus, H. L. (1980). *A five stage model of the mental activities involved in directed skill acquisition*. (Operations Research Center, Ed.). Berkeley, CA: University of California.
- Dudek N, Marks M, Lee C, W. T. (2008). Assessing the quality of supervisors’ completed clinical evaluation reports. *Medical Education*, 42(8), 816–822.
- Dudek, N., & Dojeiji, S. (2014). Twelve tips for completing quality in-training evaluation reports. *Medical Teacher*, 36, 1038–1042. <http://doi.org/10.3109/0142159X.2014.932897>
- Dudek, N. L., Marks, M. B., & Regehr, G. (2005). Failure to fail: the perspectives of clinical supervisors. *Academic Medicine : Journal of the Association of American Medical Colleges*, 80(10 Suppl), S84–7.
- Dudek, N. L., Marks, M. B., Wood, T. J., Dojeiji, S., Bandiera, G., Hatala, R., ... Sadownik, L. (2012). Quality evaluation reports: Can a faculty development program make a difference? *Medical Teacher*, 34(11), e725–31. <http://doi.org/10.3109/0142159X.2012.689444>
- Dudek, N, Papp, S, Gofton, W. (n.d.). Going paperless? Issues in converting a surgical assessment tool to an electronic version. *In Press*.
- Elstein, A. S., Shulman, L. S., & Sprafka, S.A. (1981). Medical Problem Solving. *Academic Medicine*, 56, 76-77.

- Epstein, R. M. (2007). Assessment in medical education. *The New England Journal of Medicine*, 356(20), 2109; author reply 2110. <http://doi.org/10.1056/NEJMc070466>
- Epstein, R., Hundert, E. (2002). Defining and assessing professional competence. *JAMA*, 287(2), 226–235.
- Espinell, A. G., Shah, R. K., Beach, M. C., & Boss, E. F. (2014). What parents say about their child's surgeon: parent-reported experiences with pediatric surgical physicians. *JAMA Otolaryngology-- Head & Neck Surgery*, 140(5), 397–402. <http://doi.org/10.1001/jamaoto.2014.102>
- Eva, K. W., Armson, H., Holmboe, E., Lockyer, J., Loney, E., Mann, K., & Sargeant, J. (2012). Factors influencing responsiveness to feedback: On the interplay between fear, confidence, and reasoning processes. *Advances in Health Sciences Education*, 17(1), 15–26. <http://doi.org/10.1007/s10459-011-9290-7>
- Eva, K. W., Hodges, B. D. (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical Education*, 46, 914–919.
- Fisher, C. (2008). What if we took within-person performance variability seriously? *Ind Organ Psychol*, 1, 185–189.
- Fitzgerald, J, Marron, C, G. C. (2011). The influence of specialty, grade, gender and deanery on the implementation and outcomes of European working time regulations in surgery. *British Journal of Surgery*, 98(S3), 21.
- Fitzpatrick, R., Bowling, A., Gibbons, E., Haywood, K., Jenkinson, C., Mackintosh, A, Peters, M. (2013). *A structured review of patient-reported measures in relation to selected chronic conditions, perceptions of quality of care and career impact*. Oxford.
- Fontes, R. B. V., Selden, N. R., & Byrne, R. W. (2014). Fostering and Assessing Professionalism and Communication Skills in Neurosurgical Education. *Journal of Surgical Education*, 71(6), e83–e89. <http://doi.org/10.1016/j.jsurg.2014.06.016>
- Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., De Rossi, S., & Horsley, T. (2010). Toward a definition of competency-based education in medicine: a systematic review of published definitions. *Medical Teacher*, 32(8), 631–637. <http://doi.org/10.3109/0142159X.2010.500898>
- Fuller, A., Hodkinson, H., Hodkinson, P., & Unwin, L. (2005). Learning as peripheral participation in communities of practice: a reassessment of key concepts in workplace learning. *British Educational Research Journal*, 31(1), 49–68. <http://doi.org/10.1080/0141192052000310029>
- George, B. C., Teitelbaum, E. N., Meyerson, S. L., Schuller, M. C., DaRosa, D. a., Petrusa, E. R., ... Fryer, J. P. (2014). Reliability, Validity, and Feasibility of the Zwisch Scale for the Assessment of Intraoperative Performance. *Journal of Surgical Education*, 71(6), e90–e96. <http://doi.org/10.1016/j.jsurg.2014.06.018>

- Gillespie, M. (2005). Student-teacher connection: A place of possibility. *Journal of Advanced Nursing*, 52(2), 211–219. <http://doi.org/10.1111/j.1365-2648.2005.03581.x>
- Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors. *Academic Medicine*, 86(10), S1-S7. <http://doi.org/10.1097/ACM.0b013e31822a6cf8>
- Ginsburg, S., Eva, K., & Regehr, G. (2013). Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Academic Medicine*, 88(10), 1539–44. <http://doi.org/10.1097/ACM.0b013e3182a36c3d>
- Ginsburg, S., Gold, W., Cavalcanti, R. B., Kurabi, B., & McDonald-Blumer, H. (2011). Competencies “Plus”: The Nature of Written Comments on Internal Medicine Residents’ Evaluation Forms. *Academic Medicine*, 86(10), S30–S34. <http://doi.org/10.1097/ACM.0b013e31822a6d92>
- Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Academic Medicine : Journal of the Association of American Medical Colleges*, 85(5), 780–786. <http://doi.org/10.1097/ACM.0b013e3181d73fb6>
- Glass, J. M. (2014). Competency based training is a framework for incompetence. *BMJ*, 348(apr25_11), g2909. <http://doi.org/10.1136/bmj.g2909>
- Gofton, W. T., Dudek, N. L., Wood, T. J., Balaa, F., & Hamstra, S. J. (2012). The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE). *Academic Medicine*, 87(10), 1401-1407. <http://doi.org/10.1097/ACM.0b013e3182677805>
- Goitein, L. (2015). Training Young Doctors: the Current Crisis. *The New York Review of Books*.
- Govaerts, M. J. B., Van Der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education*, 12(2), 239–260. <http://doi.org/10.1007/s10459-006-9043-1>
- Govaerts, M. J., van der Vleuten, C. P., Schuwirth, L. W., Muijtjens, A. M. (2005). The use of observational diaries in in-training evaluation: student perception. *Advances in Health Sciences Education*, 10, 171-188.
- Govaerts, M., van der Vleuten, C. (2013). Validity in work-based assessment: expanding our horizons. *Medical Education*, 47(12), 1164–1174.
- Grantcharov, T. P., & Reznick, R. K. (2009). Training tomorrow’s surgeons: What are we looking for and how can we achieve it? In *ANZ Journal of Surgery* (Vol. 79, pp. 104–107). <http://doi.org/10.1111/j.1445-2197.2008.04823.x>

- Gray, J. D. (1996). Global rating scales in residency education. *Academic Medicine : Journal of the Association of American Medical Colleges*, 71(1 Suppl), S55–S63. <http://doi.org/10.1097/00001888-199601000-00043>
- Guskey, T. R. (2014). Evaluating Professional Learning. In S. Billett, C. Harteis, & H. Gruber (Eds.), *International Handbook of Research in Professional and Practice-based Learning* (pp. 1215–1235). Dordrecht: Springer Netherlands. <http://doi.org/10.1007/978-94-017-8902-8>
- Hamstra, S. J. (2012). Keynote address: The focus on competencies and individual learner assessment as emerging themes in medical education research. *Academic Emergency Medicine*, 19(12), 1336–1343. <http://doi.org/10.1111/acem.12021>
- Hanson, J, Rosenberg, A, Lane, J. (2013). Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Frontiers in Psychology*, 4, 1–10.
- Hattie, J, Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(81), 81–112. <http://doi.org/10.3102/003465430298487>
- Hauer, K. E., ten Cate, O., Boscardin, C., Irby, D. M., Iobst, W., & O’Sullivan, P. S. (2014). Understanding trust as an essential element of trainee supervision and learning in the workplace. *Advances in Health Sciences Education*, 19(3), 435–456. <http://doi.org/10.1007/s10459-013-9474-4>
- Hodges, B. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Medical Teacher*, 35(7), 564–8. <http://doi.org/10.3109/0142159X.2013.789134>
- Hodges, B. D. (2010). A tea-steeping or i-Doc model for medical education? *Academic Medicine : Journal of the Association of American Medical Colleges*, 85(9 Suppl), S34–S44. <http://doi.org/10.1097/ACM.0b013e3181f12f32>
- Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher*, 32(8), 676–682. <http://doi.org/10.3109/0142159X.2010.500704>
- Holmboe, E, Ward, D, Reznick, R, Katsufakis, P, Leslie, K, Patel, V., & Ray, D, Nelson, E. (2011). Faculty development in assessment: The missing link in competency-based medical education. *Academic Medicine*, 86(4), 460–467.
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288. <http://doi.org/10.1177/1049732305276687>

- Humphrey-Murto, S, Varpio, L, Gonsalves, C, Wood, T.J. Twelve tips for demonstrating methodological rigour for the Delphi and Nominal Group consensus techniques in medical education research. *Manuscript in Progress*, 2015.
- Jennett, P, Scott, S, Atkinson, M. (1995). Patient charts and office management decisions: chart audit and chart stimulated recall. *Journal of Continuing Education in the Health Professions*, 15, 31–9.
- Jones, A. E., & Phillips, A. W. (2014). Response to: Reflections on competency-based education and training for surgical residents. *Journal of Surgical Education*, 71(5), 651. <http://doi.org/10.1016/j.jsurg.2014.04.007>
- Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. *BMJ*, 311(August), 376–380.
- Jones, M., van Kessel, G., Swisher, L., Beckstead, J., & Edwards, I. (2014). Cognitive maps and the structure of observed learning outcome assessment of physiotherapy students' ethical reasoning knowledge. *Assessment & Evaluation in Higher Education*, 39(1), 1–20. <http://doi.org/10.1080/02602938.2013.772951>
- Kane, M. (1992). The assessment of professional competence. *Evaluation and the Health Professions*, 15, 163–82.
- Kogan, J. R., Conforti, L. N., Bernabeo, E. C., Durning, S. J., Hauer, K. E., & Holmboe, E. S. (2012). Faculty staff perceptions of feedback to residents after direct observation of clinical skills. *Medical Education*, 46(2), 201–215. <http://doi.org/10.1111/j.1365-2923.2011.04137.x>
- Kogan, J. R., & Holmboe, E. (2013). Realizing the promise and importance of performance-based assessment. *Teaching and Learning in Medicine*, 25 Suppl 1(April 2014), S68–74. <http://doi.org/10.1080/10401334.2013.842912>
- Kogan, J, Conforti, L, Iobst, W, Holmboe, E. (2014). Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Academic Medicine*, 89(5), 721–7.
- Kogan, J, Holmboe, E, Hauer, K. (2009). Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review. *JAMA*, 302(12), 1316–1326.
- Kopelow, M, Schnable, G, Hassard, T, Tamblyn, R, Klass, D, Beazley, G, Hechter, F, Grott, M. (1992). Assessing practicing physicians in two settings using standardized patients. *Academic Medicine*, 67, S19–21.
- Korte, R. C., Beeson, M. S., Russ, C. M., Carter, W. a, & Reisdorff, E. J. (2013). The emergency medicine milestones: A validation study. *Academic Emergency Medicine*, 20(7), 730–735. <http://doi.org/10.1111/acem.12166>

- Krueger, R., Casey, M. (2014). *Focus groups: a practical guide for applied research* (5th ed.). Los Angeles, CA: Sage Publications.
- Lave, J; Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.
- Lee, A, Oetting, T, Beaver, H, Carter, K. (2009). The ACGME outcome project in ophthalmology: practical recommendations for overcoming the barriers to local implementation of the national mandate. *Survey of Ophthalmology*, 54(4), 507–17.
- Lewis, F, Klingensmith, M. (2012). Issues in general surgery residency training. *Annals of Surgery*, 256(4), 553–559.
- Lingard, L., Espin, S., Evans, C., & Hawryluck, L. (2004). The rules of the game: interprofessional collaboration on the intensive care unit team. *Critical Care (London, England)*, 8(6), R403–R408.
<http://doi.org/10.1186/cc2958>
- List, D. (2001). The Consensus Group Technique in Social Research. *Field Methods*, 13(3), 277–290.
- Lockyer, J. (2013). Multi-source feedback: can it meet criteria for good assessment? *Journal of Continuing Education in the Health Professions*, 33, 89–98.
- Lurie, S. J. (2012). History and practice of competency-based assessment. *Medical Education*, 46(1), 49–57.
<http://doi.org/10.1111/j.1365-2923.2011.04142.x>
- Lurie, S, Mooney, C, Lyness, J. (2009). Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: A systematic review. *Academic Medicine*, 84, 301–309.
- Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annual Review of Psychology*, 51, 93–120. <http://doi.org/10.1146/annurev.psych.51.1.93>
- Mann, K. V. (2011). Theoretical perspectives in medical education: Past experience and future possibilities. *Medical Education*, 45(1), 60–68. <http://doi.org/10.1111/j.1365-2923.2010.03757.x>
- Maudsley, R, Wilson, D, Neufeld, V. (2000). Education future physicians for Ontario: Phase II. *Academic Medicine*, 75(2), 113–26.
- Maxwell, A, Crocker, M, Jones, T, Bhagawati, D, Bhagawati, D, Papadopoulos, M, Bell, B. (2010). Implementation of the European working time directive in neurosurgery reduces continuity of care and training opportunities. *Acta Neurochir*, 152, 1207–10.

- McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education*, 6, 42. <http://doi.org/10.1186/1472-6920-6-42>
- Meade, L. B., Borden, S. H., Mcardle, P., Rosenblum, M. J., Picchioni, M. S., & Hinchey, K. T. (2012). From theory to actual practice: Creation and application of milestones in an internal medicine residency program, 2004–2010. *Medical Teacher*. <http://doi.org/10.3109/0142159X.2012.689441>
- Mellinger, J. D., Damewood, R., & Morris, J. B. (2015). Assessing the Quality of Graduate Surgical Training Programs: Perception vs Reality. *Journal of the American College of Surgeons*, 220(5), 785–789. <http://doi.org/10.101>
- Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer and H. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum. 6/j.jamcollsurg.2015.01.002
- Miller, G.E. (1990). The assessment of clinical skills, competence, performance. *Academic Medicine*, 64(9), S63-67.
- Mulder, H., Ten Cate, O., Daalder, R., & Berkvens, J. (2010). Building a competency-based workplace curriculum around entrustable professional activities: The case of physician assistant training. *Medical Teacher*, 32(10), e453–e459. <http://doi.org/10.3109/0142159X.2010.513719>
- Murphy, M, Black, N, Lmaping, D, McKee, C, Sanderson, C, Askham, Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development. *Health Technology Assessment*, 2(3), 1–88.
- Newton, J. M., Henderson, A., Jolly, B., & Greaves, J. (2015). A contemporary examination of workplace learning culture: An ethnomethodology study. *Nurse Education Today*, 35(1), 91–96. <http://doi.org/10.1016/j.nedt.2014.07.001>
- Noel, G, Herbers, J, Caplow, M, Cooper, G, Pangaro, L, Harvey, J. (1992). How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine*, 117, 757–65.
- Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The Mini-CEX: A Method for Assessing Clinical Skills. *Ann Intern Med*, 138(6), 476–481.
- Norcini, J, Blank, L, Arnold, G, Kimball, H. (1995). The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Annals of Internal Medicine*, 123, 795–99.
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Science Education*. Published online Feb 10, 1-8.

- Norman, G., Bordage, G., Page, G., Keane, D. (2006). How specific is case specificity? *Medical Education*, 40, 618-623.
- Opila, D. (1997). The impact of feedback to medical housestaff on chart documentation and quality of care in the outpatient setting. *J Gen Intern Med*, 12(6), 352–356.
- Pelgrim, E. A., Kramer, A. W., Mokkink, H. G., Van der Vleuten, C. P. (2012). Quality of written narrative feedback and reflection in a modified mini-clinical evaluation exercise: an observational study. *BMC Medical Education*, 12, 97-103.
- Pugh, D., Hamstra, S. J., Wood, T. J., Humphrey-Murto, S., Touchie, C., Yudkowsky, R., & Bordage, G. (2014). A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. *Advances in Health Sciences Education*, 20, 85-100. <http://doi.org/10.1007/s10459-014-9512-x>
- Purcell, J, Tarpley, J. (2009). How long does it take to train a surgeon? *BMJ*, 339, b4260.
- Rees, C. E., & Knight, L. V. (2007). The trouble with assessing students' professionalism: theoretical insights from sociocognitive psychology. *Academic Medicine : Journal of the Association of American Medical Colleges*, 82(1), 46–50. <http://doi.org/10.1097/01.ACM.0000249931.85609.05>
- Regehr G, Eva K, Ginsburg S, Halwani Y, S. R. (2011). *Assessment in Postgraduate Medical Education : Trends and Issues in Assessment in the Workplace. Members of the Future of Medical Education in Canada Postgraduate Project consortium.*
- Regehr, G., Bogo, M., Regehr, C., & Power, R. (2007). Can we build a better mousetrap? Improving the measures of practice performance in the field practicum. *Journal of Social Work Education*, 43(2), 327-344. <http://doi.org/10.5175/JSWE.2007.200600607>
- Regehr, G., MacRae, H., Reznick, R., Szalay, D. (1998). Comparing the Psychometric Properties of Checklists and Global Rating Scales. *Academic Medicine*, 73, 993–997.
- Reznick, R., Regehr, G., MacRae, H., Martin, J., & McCulloch, W. (1997). Testing technical skill via an innovative “bench station” examination. *American Journal of Surgery*, 173(3), 226–30.
- Rogoff, B. (1995). Observing sociocultural activity on three planes: participatory appropriation, guided participation, apprenticeship. In P. Wertsch, W. Alvarez, A. del Rio (Ed.), *Sociocultural Studies of Mind* (pp. 139–164). Cambridge: Cambridge University Press.
- Rogoff, B, Lave, J. (1984). *Everyday Cognition: Its Development in Social Context*. Cambridge: Harvard University Press.

- Royal College of Physicians and Surgeons of Canada. (2014). Competency by Design. Retrieved May 26, 2015, from <http://www.royalcollege.ca/portal/page/portal/rc/resources/cbme>
- Rushton, J, Hicks, P, Carraccio, C. (2010). The next phase of pediatric residency education: the partnership of the milestones project. *Academic Pediatrics*, 10(3), 91–2.
- Sackett, P, Zedeck, S, Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, 73, 482–6.
- Saleem, J, Russ, A, Justice, C, Hagg, H, Ebright, P, Woodbridge, P, Doebbeling, B. (2009). Exploring the persistence of paper with the electronic health record. *International Journal of Medical Informatics*, 78, 618–28.
- Sanfey, H. (2014). Assessment of surgical training. *The Surgeon*, 12(6), 350–356. <http://doi.org/10.1016/j.surge.2014.07.001>
- Santen, S, Rademacher, N, Heron, S, Khandelwal, S, Hauff, S, Hopson, L. (2010). How Competent Are Emergency Medicine Interns for Level 1 Milestones: Who Is Responsible? *Academic Emergency Medicine*, 20(7), 736–739.
- Sargeant, J. (2012). Qualitative Research Part II: Participants, Analysis, and Quality Assurance. *Journal of Graduate Medical Education*, 4(1), 1–3. <http://doi.org/10.4300/JGME-D-11-00307.1>
- Schuwirth, L., & Ash, J. (2013). Assessing tomorrow's learners: in competency-based education only a radically different holistic method of assessment will work. Six things we could forget. *Medical Teacher*, 35(7), 555–9. <http://doi.org/10.3109/0142159X.2013.787140>
- Schuwirth, L, Van der Vleuten, C. (2011). Programmatic assessment: from assessment of learning to assessment for learning. *Medical Teacher*, 33, 478–485.
- Sherbino, J. (2009). *A daily encounter card to assess undergraduate emergency medicine trainees*. CanMEDS Best Practice Abstract. Royal College of Physicians and Surgeons of Canada.
- Sherbino, J., Kulasegaram, K., Worster, A., & Norman, G. R. (2013). The reliability of encounter cards to assess the CanMEDS roles. *Advances in Health Sciences Education*, 18(5), 987–996. <http://doi.org/10.1007/s10459-012-9440-6>
- Sidhu, R. S., Grober, E. D., Musselman, L. J., & Reznick, R. K. (2004). Assessing competency in surgery: Where to begin? *Surgery*, 135(1), 6–20. [http://doi.org/10.1016/S0039-6060\(03\)00154-5](http://doi.org/10.1016/S0039-6060(03)00154-5)

- Sonnadara, R. R., Mui, C., McQueen, S., Mironova, P., Nousiainen, M., Safir, O., ... Reznick, R. (2013). Reflections on Competency-Based Education and Training for Surgical Residents. *Journal of Surgical Education*, 1–8. <http://doi.org/10.1016/j.jsurg.2013.06.020>
- Streiner, D, Norman, G. (2008). *Health measurement scales: a practical guide to their development* (4th ed.). Oxford, UK: Oxford University Press.
- Studnicki, J, Bradham, D, Marshburn, J, Foulis, P, Straumfjord, J. (1993). A feedback system for reducing excessive laboratory tests. *Arch Pathol Lab Med*, 117(1), 35–39.
- Sturman, M, Cashen, L, Cheramie, R. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *J Appl Psychol*, 90(2), 269–283.
- Sullivan, G, Feinn, R. (2012). Using effect size – or why the P value is not enough. *Journal of Graduate Medical Education*, September, 279–282.
- Taber, S., Frank, J. R., Harris, K. a, Glasgow, N. J., Iobst, W., & Talbot, M. (2010). Identifying the policy implications of competency-based education. *Medical Teacher*, 32(8), 687–691. <http://doi.org/10.3109/0142159X.2010.500706>
- Talbot, M. (2004). Monkey see, monkey do: a critique of the competency model in graduate medical education. *Medical Education*, 38, 587–592.
- Tammela, O. (2013, February). Applications of consensus methods in the improvement of care of paediatric patients: A step forward from a “good guess.” *Acta Paediatrica, International Journal of Paediatrics*. 102(2), 111-115. <http://doi.org/10.1111/apa.12120>
- Telio, S., Ajjawi, R., & Regehr, G. (2014). The “Educational Alliance” as a Framework for Reconceptualizing Feedback in Medical Education. *Academic Medicine*, 90(5), 609-614. <http://doi.org/10.1097/ACM.0000000000000560>
- Ten Cate, O. (2005). Entrustability of professional activities and competency-based training. *Medical Education*. 39(12), 1176-1177. <http://doi.org/10.1111/j.1365-2929.2005.02341.x>
- Ten Cate, O. (2013). Competency-based education, entrustable professional activities, and the power of language. *Journal of Graduate Medical Education*, 5(1), 6–7. <http://doi.org/10.4300/JGME-D-12-00381.1>
- Ten Cate, O., & Scheele, F. (2007). Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Academic Medicine : Journal of the Association of American Medical Colleges*, 82(6), 542–547. <http://doi.org/10.1097/ACM.0b013e31805559c7>

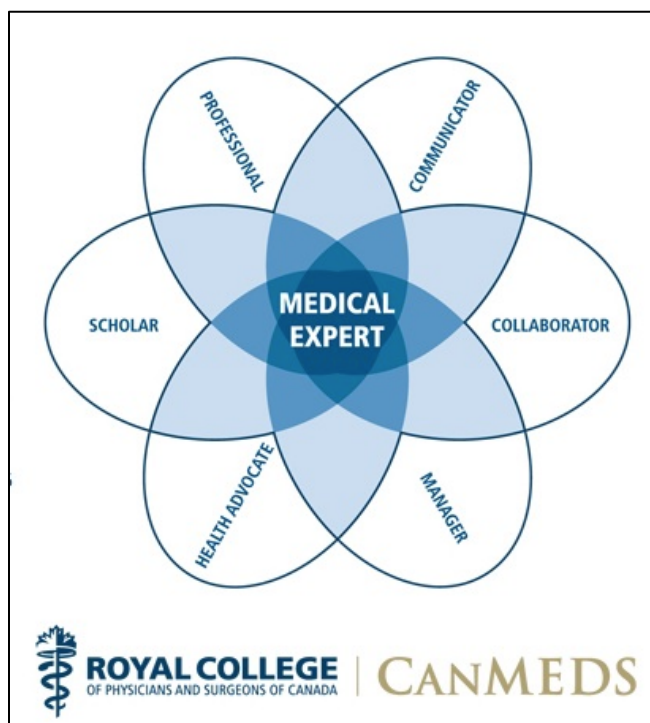
- Tesser, A. R. S. (1975). The reluctance to transmit bad news. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (3rd ed., pp. 193–232). New York: Academic.
- Touchie, C. (2013). Conference proceeding from CUSEC (Canadian Undergraduate Surgical Education Committee) Meeting. In *CUSEC (Canadian Undergraduate Surgical Education Committee) Meeting*. Ottawa, Canada.
- Turnbull, J., Van Barneveld, C. (2002). Assessment of clinical performance: In-training evaluation. In D. Norman, G. Van der Vleuten, C. Newble (Ed.), *International handbook of medical education* (pp. 793–810). Dordrecht: Kluwer.
- Van de Ridder, J, Stokking K, McGaghie, W, ten Cate, O. (2008). What is feedback in clinical education? *Medical Education*, 42(2), 189–197. <http://doi.org/10.1111/j.1365-2923.2007.02973.x>.
- Van der Vleuten, C., Schuwirth, L. (2005). Assessment of professional competence: from methods to programmes. *Medical Education*, 39, 309–17.
- Van der Vleuten, C., & Verhoeven, B. (2013). In-training assessment developments in postgraduate education in Europe. *ANZ Journal of Surgery*, 83(6), 454–459. <http://doi.org/10.1111/ans.12190>
- Van der Vleuten, C, Schuwirth, L, Scheele, F, Driessen, E, Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Pract Res Cl Ob*, 24(6), 703–719.
- Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., Van Tartwijk, J. (2012). A model of programmatic assessment fit for purpose. *Medical Teacher*, 34, 205–214.
- Van Eaton, E. G., Tarpley, J. L., Solorzano, C. C., Cho, C. S., Weber, S. M., & Termuhlen, P. M. (2011). Resident education in 2011: three key challenges on the road ahead. *Surgery*, 149(4), 465–73. <http://doi.org/10.1016/j.surg.2010.11.007>
- Veloski, J, Boex, JR, Grasberger, MJ, Evans, A, Wolfson, D. (2006). Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide No. 7. *Medical Teacher*, 28(2), 117–128.
- Wagner, J. P., Chen, D. C., Donahue, T. R., Quach, C., Joe Hines, O., Hiatt, J. R., & Tillou, A. (2014). Assessment of Resident Operative Performance Using a Real-Time Mobile Web System: Preparing for the Milestone Age. *Journal of Surgical Education*, pp. 1–6. Elsevier. <http://doi.org/10.1016/j.jsurg.2014.06.008>
- Warm, E. J., Mathis, B. R., Held, J. D., Pai, S., Tolentino, J., Ashbrook, L., ... Mueller, C. (2014). Entrustment and mapping of observable practice activities for resident assessment. *Journal of General Internal Medicine*. 29(8), 1177–1182. <http://doi.org/10.1007/s11606-014-2801-5>

- Weller, J. M., Jones, a., Merry, a. F., Jolly, B., & Saunders, D. (2009). Investigation of trainee and specialist reactions to the mini-Clinical Evaluation Exercise in anaesthesia: Implications for implementation. *British Journal of Anaesthesia*, 103(4), 524–530. <http://doi.org/10.1093/bja/aep211>
- Wenger, E. (1998). *Communities of Practice. Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.
- Wigton, R. (1980). The effects of student personal characteristics on the evaluation of clinical performance. *Journal of Medical Education*, 55, 423–7.
- Wigton, R., Patil, K., Hoellerich, V. (1986). The effect of feedback in learning clinical diagnosis. *Journal of Medical Education*, 61(10), 816–822.
- Wilkinson, J. R., Crossley, J. G. M., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education*, 42(4), 364–373. <http://doi.org/10.1111/j.1365-2923.2008.03010.x>
- Williams, R. G., Klamen, D. a, & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine*, 15(4), 270–292. http://doi.org/10.1207/S15328015TLM1504_11
- Wilson, A. B., Choi, J. N., Torbeck, L. J., Mellinger, J. D., Dunnington, G. L., & Williams, R. G. (2014). Clinical Assessment and Management Examination-Outpatient (CAMEO): Its Validity and Use in a Surgical Milestones Paradigm. *Journal of Surgical Education*, pp. 1–8. Elsevier. <http://doi.org/10.1016/j.jsurg.2014.06.010>
- Woehr, D., Huffcutt, A. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67, 189–205.
- Wood, T. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education*, 19(3), 409–427.
- Wright, E., Langille, M. (2011). Implementation and experience with a standardized assessment of a clinical encounter report (STACER) in a surgical residency program. *Open Medicine*, 5(3), available online.
- Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013). Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education*, 18(3), 325–341. <http://doi.org/10.1007/s10459-012-9372-1>
- Yeates, P, O'Neill, P, Mann, K, & Eva, K. (2013). “You’re certainly relatively competent’: assessor bias due to recent experiences. *Medical Education*, 47(9), 910–922.

Zibrowski, E, Singh, S, Goldszmidt, M, Watling, C, Kenyon, C, & S. V. (2009). The sum of the parts detracts from the intended whole: Competencies and in-training assessments. *Medical Education*, 43, 741–748.

Appendices

Appendix A: The CanMEDS Roles



The CanMEDS Roles

Appendix B: Comparison of the CanMEDS and ACGME Competency Frameworks*

Roles Frameworks: Canada and the United States

EFPO Roles (approved 1992) *

- Medical expert, clinical decision maker
- Communicator, educator, humanist, healer.
- Collaborator
- Gatekeeper, resource manager
- Learner
- Scientist, scholar
- Health advocate
- Person

RCPSC/CanMEDS (approved 1996) *

- Medical expert
- Communicator
- Collaborator
- Manager
- Scholar
- Health advocate
- Professional

MSOP Attributes (approved 1998) *

- Knowledgeable
- Skillful
- Dutiful
- Altruistic

ACGME/Core Competencies (approved 1999) *

- Patient care
- Medical knowledge
- Interpersonal and communication skills
- Systems-based practice
- Practice-based learning and improvement
- Professionalism

* EFPO stands for Educating Future Physicians for Ontario; RCPSC stands for Royal College of Physicians and Surgeons of Canada; MSOP stands for Medical School Objectives Project; and ACGME stands for the Accreditation Council for Graduate Medical Education.

* Taken from Hodges, 2010

Appendix C: The Zwisch Scale*

Level	Descriptor
1	Show and Tell
2	Active Help
3	Passive Help
4	Supervision Only

* Taken from George et al, 2013

Appendix D: The Ottawa Surgical Competency Operating Room (O-SCORE) scale**

Level	Descriptor
1	"I had to do" <i>i.e., Requires complete hands on guidance, did not do, or was not given the opportunity to do</i>
2	"I had to talk them through" <i>i.e., Able to perform tasks but requires constant direction</i>
3	"I had to prompt them from time to time" <i>i.e., Demonstrates some independence, but requires intermittent direction</i>
4	"I needed to be there in the room just in case" <i>i.e., Independence but unaware of risks and still requires supervision for safe practice</i>
5	"I did not need to be there" <i>i.e., Complete independence, understands risks and performs safely, practice ready</i>

** Taken from Gofton, et al 2012

Appendix E: CBME Challenges for Surgery & Potential Solutions

CHALLENGE	POTENTIAL SOLUTIONS
Achieving sufficient surgical exposure considering increased number of conflicting lectures, CanMEDS teaching and simulation sessions and decreased surgical volume in the Academic Centers and greater number of residents	- Simulation lab used for deliberate practice - Create larger role for community hospitals - University affiliation for community surgeons for maintenance of skills, as well as teaching role recognition - Ownership of education taken by residents to seek out training in areas of weakness - Increased admin personnel to guide residents to ORs based on milestone achievement * - Removal of extraneous rotations and “scut work”/unhelpful elements of service once resident has mastered them to streamline surgical exposure
Assessment & feedback delivery require improvement both in quality and frequency, as well as culture change so surgeons give more daily formative feedback and residents get used to hearing constructive criticism	- Developing assessment tools which are valid, reliable, and predictive of future success is the immediate challenge - Collect validity evidence on each assessment tool in use, specifically in surgical content - Create feasible assessment tools without jargon, targeted to surgeon raters - Multiple, short, quality assessments to be completed by multiple staff (less time-consuming) - Staff surgeon buy-in required - Formal training or orientation sessions to improve feedback and to increase proper use of assessment tools - Increase remuneration of teaching staff for education, raising its status to the equivalent of clinical work - Create culture of constructive criticism within residency programs - Disseminate knowledge of modern validity theory
Potential reductionism of Milestones and EPAs (tendency to complete the checklist without surgical resident judgment and decision making actually advancing)	- Global Assessment Items in rating scales to capture subjective gestalt of surgeon supervisors - Incorporate qualitative assessment into checklist-dominated culture - Maintain a resident assessment committee which meets at least quarterly, made up of multiple members, who review resident’s portfolio (program should not rely on just one person reading checklist assessment forms)
Defining & assessing non-technical (aka soft or intrinsic) skills	- Holistic milestones based on CanMEDS non-medical expert roles - Write EPAs that supersede cultural and contextual boundaries (multiple stakeholder opinions) - Disseminate knowledge of domains encompassed by roles so that surgeons know what they are assessing
Service to education ratio; Resident service is required in our system	- Increase staff surgeon ward expectations** - Change funding model to fill service roll with alternative service model professionals (i.e. Nurse Practitioners, Physician Assistants), task-shifting or task-sharing
Maintenance of competence (once	- Routinely check maintenance of competence in sim lab

milestones have been achieved in a particular area)	based assessment - Require residents to record ORs on call - Establish baseline # of procedures to achieve & maintain competence - Electronic logbook, linked to assessment tools, to show resident progression through milestones in real-time (self-populating database)
Incorporating the unpredictability of the Acute Care Service (ACS) into CBE	- Realistic minimum objectives/milestones set, if resident doesn't achieve them, will need to do more ACS blocks - In Senior years, cases can be logged to prove maintenance of competence achieved in junior years - Use hybrid model of CBE/traditional system, with ACS block used as a mandatory, time-based rotation (requires reliable senior resident coverage) - (In a pure CBE model, the Acute Care Model may not be compatible and would likely be dissolved)
Funding \$\$\$\$ <i>Potential Expenditures:</i> - Curriculum development start-up costs - Faculty training - Paying faculty for their involvement in feedback/assessment - Development of assessment tools - Increased #s of administrative assistants to track residents' progress and assign ORs - Research and development related to CBE - Simulation lab expenditures	Options: - Train faculty to improve assessments and feedback skills without increased remuneration (based on purely altruistic model) - Pay faculty less for clinical time and use extra \$ to increase teaching remuneration - Encourage governmental increase in funding for education-related tasks - Decrease or abolish resident's salaries as motivation to reach competence faster - Increase frequency and \$\$ of external grants from private donors for R&D (curriculum development) - Admin responsibility on resident shoulders, do not hire more admin assistants - Assume re-equilibrium of system once residents begin moving through training program faster
Resident Failure/Remediation	- Early identification of struggling residents should be more possible with regular formative feedback inherent in CBE - Formal remediation curriculum, unique to each resident, developed to address failure to achieve each of the milestones - Set maximum amounts of time a resident can be in a particular module (basis for failure out of program)
Lack of CBE Evidence Base in Surgical Literature	- Need projects comparing CBE trial programs to traditionally trained residents showing benefit - Clarify objectives, EPAs, and milestones for individual surgical divisions (CanMEDS & Medical Expert) - Seek stakeholder opinion to improve implementation (ex. ask resident opinion)

* This may not be feasible, but this is what has been done in the Toronto CBME program (Sonnadara et al, 2014)

** May also not be feasible, but alternative is likely task shifting to Physician Assistant or a similar role such as Nurse Practitioners

Appendix F: Summary of Validity Evidence Gathered

Type of Validity Evidence	OCAT
Content	- Decision to make tool mainly formative in nature - Consensus groups of experts using the nominal group process used to generate a list of assessment items and discuss the rating scale - Experts drawn from all surgical divisions including resident representation - Pre-written script prepared prior to consensus group - National email survey of experts - Pre-pilot test by experts from consensus group to screen for wording/technical issues - Assessment for content completeness in feedback sessions
Internal Structure	- Reliability data (internal consistency, generalizability analysis, group differences) - Minimum sample size calculated
Response Process	- Comparison of PGY1 data to PGY5 data - No gold standard in literature
Relations to Other Variables	- Development of items by surgeon-users in consensus group (understandable, no jargon) - Brief rater training sessions and thorough instructions present at top of tool - Logistical: paper format, attention to data security measures, and adequate scoring on entry of data - Use of entrustability scale - Global assessment item to capture subjectivity of raters - Asking intention questions in feedback sessions to hear thought process of surgeons and residents - Feasibility: Involvement of Department of Surgery, TOH Department of Surgery grant received, keep surgeon stakeholders involved throughout, cost & time to administer are low, feasibility checked in feedback groups
Consequential	- Effect on formative feedback - Effect of receiving low scores on residents - Remediation effects

Appendix G: Initial REB Approval Letter



Ottawa Health Science Network Research Ethics Board/Réseau des sciences de la santé d'Ottawa Conseil d'éthique de la recherche
Civic Box 411 725 Parkdale Avenue, Ottawa, Ontario K1Y 4E9 613-798-5555 ext. 14902 Fax: 613-761-4311
<http://www.ohri.ca/chsn-reb>

March 26, 2014

Dr. Wade Gofton
The Ottawa Hospital - Civic Campus
Division of Orthopaedic Surgery
1053 Carling Avenue, Rm J153
Ottawa, ON
K1Y 4E9

Dear Dr. Gofton:

Re: Protocol # 20130877-01H Development of an Assessment Tool for CanMEDS Competencies Observed During Surgical Clinic
Protocol approval valid until - March 25, 2015

I am pleased to inform you that this protocol underwent delegated review by the Ottawa Health Science Network Research Ethics Board (OHSN-REB) and is approved for the recruitment of English speaking participants only. No changes, amendments or addenda may be made to the protocol or the consent forms without the OHSN-REB's review and approval.

Approval is for the following:

- Protocol, version dated February 20, 2014
- Phase 1 & 2, Resident & Staff Oral Announcement, version 1, dated 14/01/20
- Phase 1 & 2, Resident & Staff Recruitment Email, version 1, dated 14/01/20
- Participant Informed Consent Form (Phase 1), version 1, dated January 20, 2014
- Participant Informed Consent Form (Phase 2), version 1, dated January 20, 2014
- Phase 1 & 2, Resident & Staff Reminder Email, version 1, dated 14/01/20
- Pre-Survey Resident Questionnaire, version 1, dated 13/12/01
- Post Survey Resident & Staff Questionnaire, version 1, dated 13/12/01

The REB no longer requires a 'valid until' date at the bottom of all approved informed consent forms. The consent forms currently approved for use by the REB are listed above.

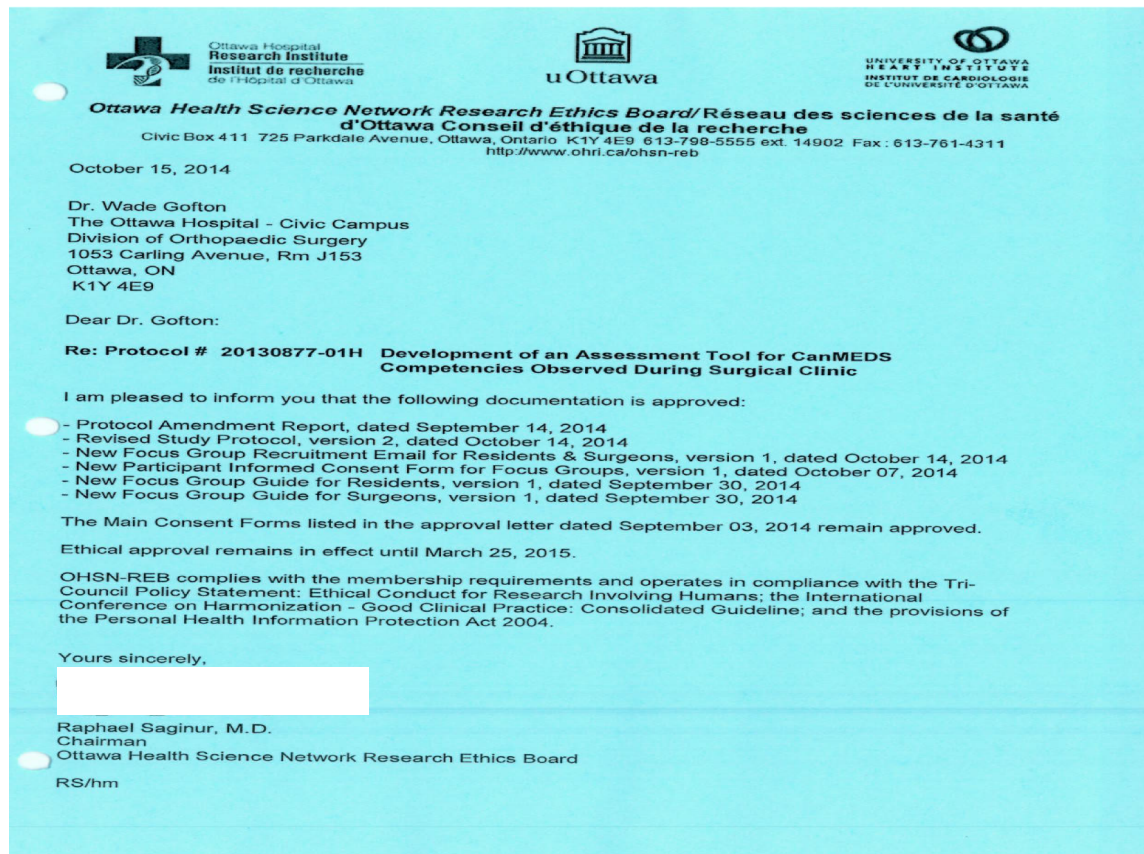
If the study is to continue beyond the expiry date noted above, a Renewal Form should be submitted to the REB approximately six weeks prior to the current expiry date. If the study has been completed by this date, a Termination Report should be submitted.

The Ottawa Health Science Network Research Ethics Board (OHSN-REB) was created by the merger of both the Ottawa Hospital Research Ethics Board (OHREB) and the Human Research Ethics Board (HREB) for meetings held at the University of Ottawa Heart Institute.

OHSN-REB complies with the membership requirements and operates in compliance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans; the International Conference on Harmonization - Good Clinical Practice: Consolidated Guideline and the provisions of the Personal Health Information Protection Act 2004.

Chairperson
Ottawa Health Science Network Research Ethics Board
/hm

Appendix H: REB Addendum Approval Letter



Appendix I: National Expert Survey Email Request

Dear Dr. _____,

I'm a PGY-3 General Surgery in Ottawa doing my Masters of Education through the University of Ottawa under the supervision of Drs. Wade Gofton and Stan Hamstra. They've identified you as an expert in the field of medical education and I was hoping to ask for your very quick help with a question related to my thesis project.

Your resident is in clinic with you today and you have been asked to assess them. What are the top 5 things you would consider to assess regarding their competence to run this clinic?

For example:

- History taking skills
 - General attitude in clinic
 - Appropriately triaging time with patients (new consults vs. follow-ups)
-

You can either email me your top 5 (plus more if you come up with them) or follow the survey monkey link provided:

<https://www.surveymonkey.com/s/MWDFRX6>

If you have a surgical colleague with an interest in medical education and feel providing their email or forwarding this to them that would be appreciated.

Thank you for your time!

Janelle Rekman

PGY-3 General Surgery
The Ottawa Hospital
M.A.Ed Candidate 2015, U of Ottawa

Appendix J: Anchor Scale Comparison: O-SCORE to OCAT

O-Score Anchor Scale

Scale

- 1—"I had to do"—i.e., *Requires complete hands on guidance, did not do, or was not given the opportunity to do*
- 2—"I had to talk them through"—i.e., *Able to perform tasks but requires constant direction*
- 3—"I had to prompt them from time to time"—i.e., *Demonstrates some independence, but requires intermittent direction*
- 4—"I needed to be in the room just in case"—i.e., *Independence but unaware of risks and still requires supervision for safe practice*
- 5—"I did not need to be there"—i.e., *Complete independence, understands risks and performs safely, practice ready*

* Taken from Gofton, W., Dudek, N., Wood, T., Balaa, F., & Hamstra, S. (2012).

OCAT Anchor Scale

Scale

- 1—"I had to do"—i.e., *Requires complete guidance, unprepared to do, or had to do for them*
- 2—"I had to talk them through"—i.e., *Able to perform but requires repeated direction*
- 3—"I had to direct them from time to time"—i.e., *Demonstrates some independence, but requires intermittent prompting*
- 4—"I needed to be available just in case"—i.e., *Independence but needs assistance with nuances of certain patients and/or situations, unable to manage all patients, still requires supervision for safe practice*
- 5—"I did not need to be there"—i.e., *Complete independence, can safely and manage a general clinic in your specialty*
- 6—"Functioning at level higher than a generalist in my specialty"—i.e., *Fellow level or subspecialty practitioner* _____

Appendix K: Resident Recruitment Email

You are being asked to participate in a research project intended to improve assessment of resident performance in the surgical clinic environment. The objective of the study is to validate a novel assessment tool which is designed to evaluate a resident's ability to function independently in surgical clinic. This assessment tool has not been used before. The participants in this first phase of the study will be general surgery and orthopaedic surgery residents at all levels at the University of Ottawa.

You must be able to read and understand English in order to participate in this research study.

Participation in this study is voluntary, and choosing not to participate will not impact your status as a surgical resident at the University of Ottawa. The authors believe participating in this study will enable you to better understand your performance in clinic due to an increased volume of direct feedback from staff surgeons, and thus may also indirectly benefit you later on during the course of your residency.

If you choose to participate in this study, we ask that at the end of each clinic day you ask the staff surgeon to take a few minutes to complete an evaluation of your performance that day. It is possible that not all the staff surgeons will have consented to participation in the study. You should ask them whether they are enrolled in the study prior to requesting an evaluation. This evaluation form will be used for study purposes only (to validate the assessment tool), and WILL NOT contribute to your program evaluation in any way. Your program director WILL NOT be given this assessment tool to use for formal evaluation. In no way will your participation in this study affect your performance evaluation for the block you're participating in. You will be asked to complete a short pre-survey at the time of consent, and a short post-survey will be sent to you by email after the study is complete.

All information you contribute to this study will be kept confidential. An independent study number will be assigned to each participant, and all research-related records will be kept in a locked office at the Ottawa Hospital. No identifying information will leave the hospital. You will not be identified in any publications or presentations that result from this study.

Please feel free to ask any questions you may have about this study. Thank you.

Further questions can be addressed to the study coordinator at: jrekman@toh.on.ca

Appendix L: Staff Surgeon Recruitment Email

You are being asked to participate in a research project intended to improve assessment of resident performance in the surgical clinic environment. The objective of the study is to validate a novel assessment tool which is designed to evaluate a resident's ability to function independently in surgical clinic. This assessment tool has not been used before but is modelled after the O-score technical skills assessment. The participants in this first phase of the study will be general surgery and orthopaedic surgery residents and staff at all levels at the University of Ottawa.

You must be able to read and understand English in order to participate in this research study.

Participation in this study is voluntary, and choosing not to participate will not impact your status as a staff surgeon at the University of Ottawa. We estimate filling out the form for a resident at the end of a day of clinic will take approximately 5 minutes, but otherwise should not add any extra load to your already busy schedule. It will be the resident's responsibility to make sure the form is submitted to the appropriate location.

If you choose to participate in this study, we ask that at the end of each clinic day you fill out an evaluation form for residents who have worked with you in clinic. Only the residents who have consented to participation in the study will request an evaluation from you. The results of the evaluation will be kept confidential. However it would be beneficial for the formative learning of the resident if you review their results with them on that day of clinic. This should be an evaluation of their performance just on that day. The resident's performance should not be rated against their peers, but instead against how ready you think they are to run the clinic independently. This evaluation form will be used for study purposes only (to validate the assessment tool), and will not contribute to program evaluation in any way during this first phase.

All information you contribute to this study will be kept confidential. An independent study number will be assigned to each participant, and all research-related records will be kept in a locked office at the Ottawa Hospital. No identifying information will leave the hospital. You will not be identified in any publications or presentations that result from this study.

Please feel free to ask any questions you may have about this study. Thank you.

Further questions can be addressed to the study coordinator at: jrekman@toh.on.ca

Appendix M: Consent Form

PARTICIPANT INFORMED CONSENT FORM

Title of Study: Development of an Assessment Tool for CanMEDS Competencies Observed During Surgical Clinic

Local Site Principal Investigator (PI): *Dr. Janelle Rekman (613-698-9494)*

Participation in this study is voluntary. Please read this Participant Informed Consent Form carefully before you decide if you would like to participate. Ask the study doctor and study team as many questions as you like. We encourage you to discuss your options with family, friends or your healthcare team.

Why am I being given this form?

You are being asked to participate in this research study because you are either a surgical resident or staff surgeon in the Department of Surgery at the Ottawa Hospital, and an assessment tool for use in surgical clinic is in development.

Why is this study being done?

With the imminent approach of competency-based surgical education, it is essential that we as surgeons are able to accurately gauge the technical and non-technical skills (CanMEDS roles) of our trainees to ensure readiness to practice independently at the completion of training. Currently, there is no appropriate assessment tool available which accurately represents a learner's true progression towards independence in the surgical clinic. We are piloting this tool at the Ottawa Hospital in the hopes that it will provide material for surgeon feedback and resident education regarding their areas of strength and weakness.

We estimate that 100 participants will be enrolled in the study from the Ottawa Hospital Department of Surgery.

How is the study designed?

There will be 2 phases of this study in which the assessment tool will be available for use in surgical clinics. Phase 1 is a 3 month pilot phase in General Surgery and Orthopedic Surgery to trial the assessment tool. Phase 2 is a 6 month full-scale trial of the assessment tool across all surgical divisions at the Ottawa Hospital.

What is expected of me?

If you consent to participate in this study, we would ask surgical residents to request completion of an assessment form at the end of each day in clinic by their staff surgeon. Staff surgeons are requested to consider the abilities of the applicable resident with respect to their ability to practice independently (not relative to their peers). Residents will be asked to return the form to one of the sealed envelopes provided, and insert it back in the study envelope for once-weekly pick-up by the study research assistant. For each completed evaluation form, resident participants will receive a \$10 Second Cup gift certificate.

It is important for both residents and surgeons to understand that these assessment forms will at this time only be used for study purposes, they will in no way reflect on a resident's performance. The assessment forms will be destroyed once data gathering is complete and they will not be returned to the program directors.

Will my research data be used in future research? No future use of this data is planned.

How long will I be involved in the study? The entire study will last approximately 3 months during Phase 1, and 6 months during Phase 2. Your participation in the study will last approximately 3 months during Phase 1, and 6 months during Phase 2.

What are the potential risks I may experience? There are no risks foreseen by participating in this study, as the evaluation forms will not contribute to summative resident program evaluation.

Questionnaires: You might not like all of the questions that you are asked. You do not have to answer any questions that make you uncomfortable.

Can I expect to benefit from participating in this research study? You may not receive any direct benefit from your participation in this study. Your participation may allow the researchers to validate this assessment tool for use in formative evaluation in surgical clinic. This may benefit future residents.

Do I have to participate? What alternatives do I have? Your participation in this study is voluntary. You may decide not to be in this study, or to be in the study now, and then change your mind later. Your decision will not affect your academic standings, current or future placements, or any future employment or potential for employment at The Ottawa Hospital.

How is my personal information being protected?

All personal health information (PHI) and your personal identifying information (PII), such as your name, address, date of birth, etc. will be kept confidential.

Release of your PHI/PII information will only be allowed if it is legally required.

As a participant, you will be assigned a coded study number that will be used throughout the study on all your study records.

A Master List provides the link between your identifying information and the coded study number.

This list will only be available to Dr. Rekman and her staff and will not leave this site.

The Master List and coded study records will be stored securely.

You will not be identified in any publications or presentations resulting from this study.

Research records will be kept for 10 years, as required by the OHSN-REB

At the end of the storage time, all paper records will be shredded and all electronic records will be securely deleted.

There are no conflicts of interest to declare related to this study.

What are my responsibilities as a study participant?

It is important to remember the following things during this study:

- Surgical residents are requested to complete an assessment form at the end of each day in clinic and then seal these forms in an envelope and return to the large study envelope.
- Staff surgeons are requested to complete an assessment form at the end of a day of clinic for each resident working with them in clinic that day.
- These evaluation forms will not be used for evaluative purposes, and will be destroyed after the study.
- Ask your Principle Investigator if you have any questions or concerns.

Will I be informed about any new information that might affect my decision to continue participating?

You will be told in a timely fashion of any new findings during the study that could affect your willingness to continue in the study. You may be asked to sign a new consent form.

Who do I contact if I have any further questions?

If you have any questions about this study, or if you feel that you have experienced a study-related injury or illness, please contact Dr. J. Rekman at 613-698-9494 or the study Research Assistant Christine Seabrook at 613-798-5555 ext 15929.

The Ottawa Health Science Network Research Ethics Board (OHSN-REB) has reviewed this protocol. The Board considers the ethical aspects of all research studies involving human participants at the The Ottawa Hospital. If you have any questions about your rights as a study participant, you may contact the Chairperson at 613-798-5555, extension 16719.

**Consent to Participate in Research**

- I understand that I am being asked to participate in a research study developing a Surgical Clinic Assessment Scale.
- This study was explained to me by _____.
- I have read, or have had it read to me, each page of this Participant Informed Consent Form.
- All of my questions have been answered to my satisfaction.
- If I decide later that I would like to withdraw my participation and/or consent from the study, I can do so at any time.
- I voluntarily agree to participate in this study.
- I will be given a copy of this signed Participant Informed Consent Form.

Participant's Printed Name	Participant's Signature	Date
----------------------------	-------------------------	------

Investigator or Delegate Statement

I have carefully explained the study to the study participant. To the best of my knowledge, the participant understands the nature, demands, risks and benefits involved in taking part in this study.

Investigator/Delegate's Printed Name	Investigator/Delegate's Signature	Date
--------------------------------------	-----------------------------------	------

Appendix N: Pre-Pilot Resident Survey

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I receive appropriate feedback regarding my performance in clinic at the end of each clinic I participate in.	1	2	3	4	5
<i>Comments</i>					

Appendix O: OCAT for Pilot Project (with Vertical Line)

The Ottawa Clinic Assessment Tool (OCAT)

Study ID: (last 4 digits of phone #)	Level: 1 2 3 4 5	Staff:
Clinic:	Date:	

The purpose of this scale is to assess the trainee's ability to safely and independently run a CLINIC IN YOUR SPECIALTY (i.e. Urology, General Surgery, etc.) at the level of a generalist (i.e. not subspecialty trained).

With that in mind please use the scale below to rate each item, irrespective of the resident's level of training. Base your ratings on the trainee's performance across the ENTIRE CLINIC (i.e. do not base your rating on only one specific patient encounter). Please complete the assessment *IMMEDIATELY* following completion of the clinic.

Scale

1—"I had to do"—i.e., Requires complete guidance, unprepared to do, or had to do for them

2—"I had to talk them through"—i.e., Able to perform but requires repeated direction

3—"I had to direct them from time to time"—i.e., Demonstrates some independence, but requires intermittent prompting

4—"I needed to be available just in case"—i.e., Independence but needs assistance with nuances of certain patients and/or situations, *unable to manage all patients*, still requires supervision for safe practice

5—"I did not need to be there"—i.e., Complete independence, can safely and manage a general clinic in your specialty

6—"Functioning at level higher than a generalist in my specialty"—i.e., Fellow level or subspecialty practitioner

						At Fellow Level
1. Clinical Preparedness	1	2	3	4	5	6
Arrived at clinic with appropriate knowledge base for anticipated patients						
2. History	1	2	3	4	5	6
Efficient data gathering						
3. Physical Exam	1	2	3	4	5	6
Efficient and accurate examination						
4. Case Presentation	1	2	3	4	5	6
Synthesis of history and physical, clear presentation						
5. Differential Diagnosis	1	2	3	4	5	6
Able to make a diagnosis and appropriately consider alternatives						
6. Management Plan	1	2	3	4	5	6
Able to develop relevant plan dependent on context and be decisive (i.e. appropriate investigations, procedures, etc)						
7. Patient/Family Communication	1	2	3	4	5	6
Effective, Sensitive, and respectful communication skills (verbal & non-verbal), language appropriate to patient understanding, able to build rapport and trust						
8. Documentation within Clinic	1	2	3	4	5	6
Orders, prescriptions, forms, etc (may not include consultation report)						

9. Collaboration Works well with and/or teaches other team members as appropriate (<i>ie staff, student, other healthcare professional</i>)	1	2	3	4	5	6
10. Time Management Able to economize time, manage interruptions, and modify time spent with individual patients appropriately	1	2	3	4	5	6
11. Attitude Engagement and ownership in running clinic	1	2	3	4	5	6
If procedures were performed in clinic:						
12. Technical Safely and effectively performs appropriate clinical procedures	1	2	3	4	5	6
13. Situational Awareness Non-technical aspects of procedure (i.e. insight into patient experience, respects patient comfort)	1	2	3	4	5	6

Resident is safe to independently manage a general clinic in your specialty

Y

N

14. Give at least 1 specific aspect of clinic done well

15. Give at least 1 specific suggestion for improvement

Staff Signature:

Appendix P: Coding Strategy for Short Answer Questions

Question #14 (*Give at least one specific aspect of clinic done well*)

1. How many of the comments contain a word or theme taken directly from numbers 1-13 on the OCAT form?
2. Code all the rest as “other”
3. Within the “other” category:
 - a. How many are completely general, supportive statements, i.e. good job
 - b. How many relate to a *very specific aspect* of clinic?
 - i. How many of these very specific ones relate to a technical skill?
 - ii. Record these statements in a separate box to discuss at the next meeting
 - c. Keep track of which “others” don’t fall into any of these categories

Question #15 (*Give at least one suggestion for improvement*)

1. How many of the comments contain a word or theme taken directly from numbers 1-13 on the OCAT form?
2. Code all the rest as “other”
3. Within the “other” category:
 - a. How many are actually a veiled compliment and not actually a suggestion for improvement? (record these in a separate location for review)
 - b. How many say: “read more”, or some variation of this.
 - c. How many relate to a *very specific aspect* of clinic?
 - i. How many of these very specific ones relate to a technical skill?
 - ii. Record these statements in a separate box to discuss at the next meeting
 - d. Keep track of which “others” don’t fall into any of these categories

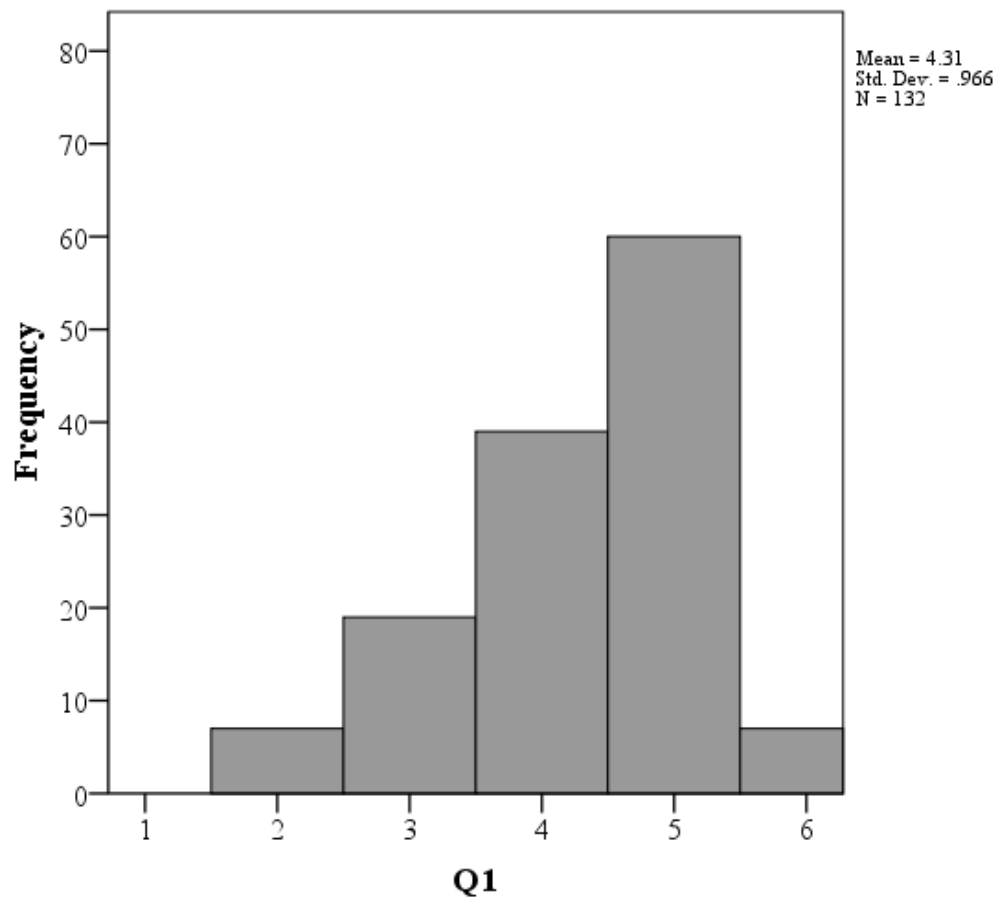
Appendix Q: Feedback Survey and Feedback Session Questions

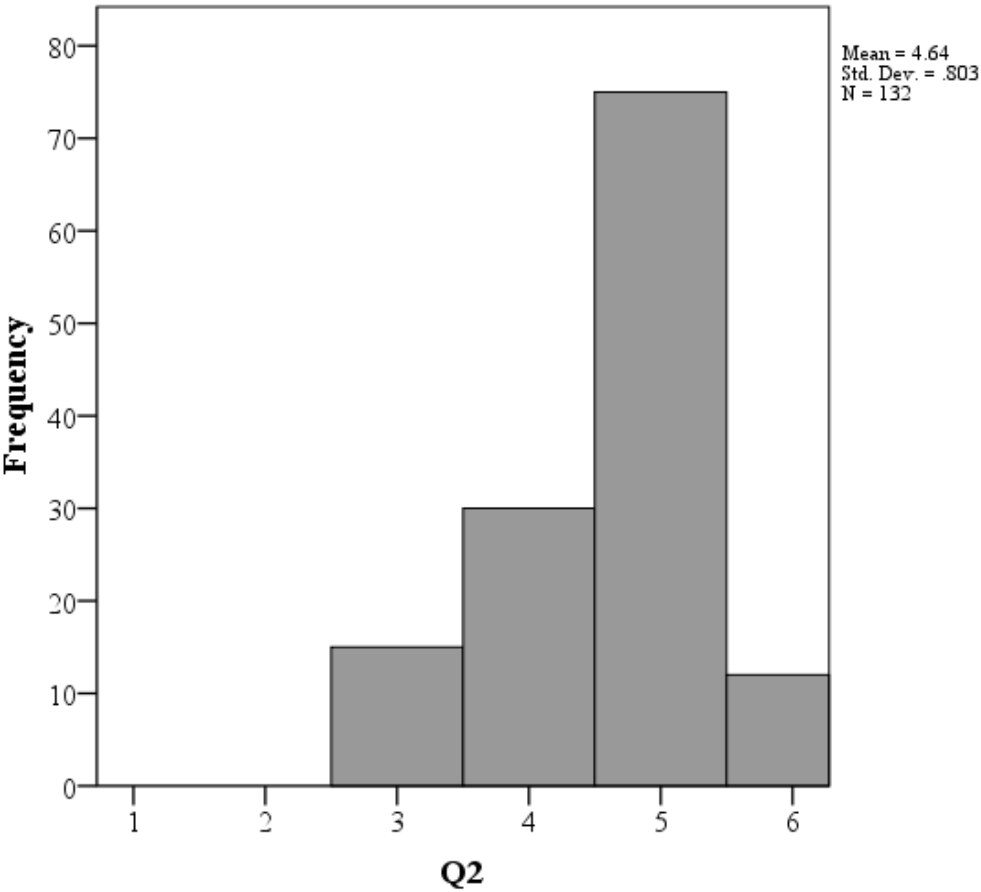
Staff:

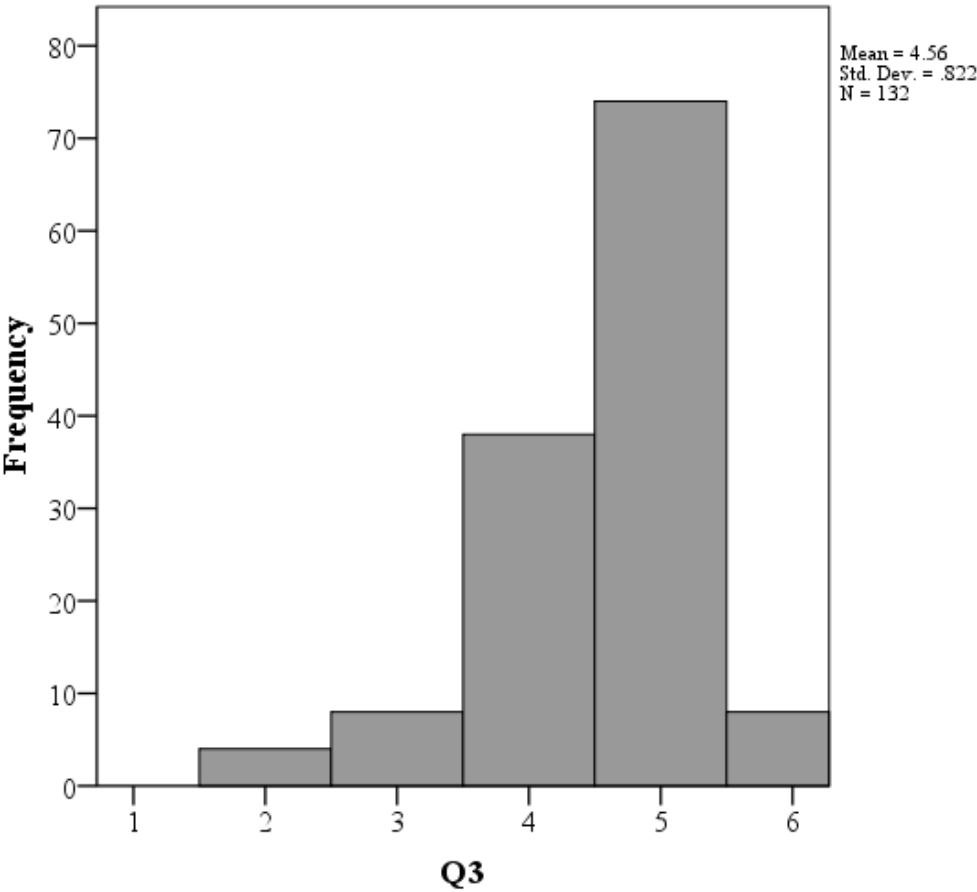
- 1) What did you think of item 1 – clinic preparedness?
- 2) What did you think of item 11 – attitude?
- 3) When working in a sub-specialized clinic (ex. uro-gynaecy, hepatobiliary, spine, etc) how did you use the rating “5”?
- 4) Did you struggle with the yes/no question? Why or why not?
- 5) What did you think of the rating “6”? How did you use it?
- 6) Overall, was the OCAT user friendly or not? Why?
- 7) How feasible is it to do an OCAT at the end of the clinic?
- 8) Did completing an OCAT impact on your verbal feedback to the resident?
- 9) Is there anything you would change about the OCAT?
- 10) Would you prefer to do the OCAT on paper or electronically on your iPad? Why?

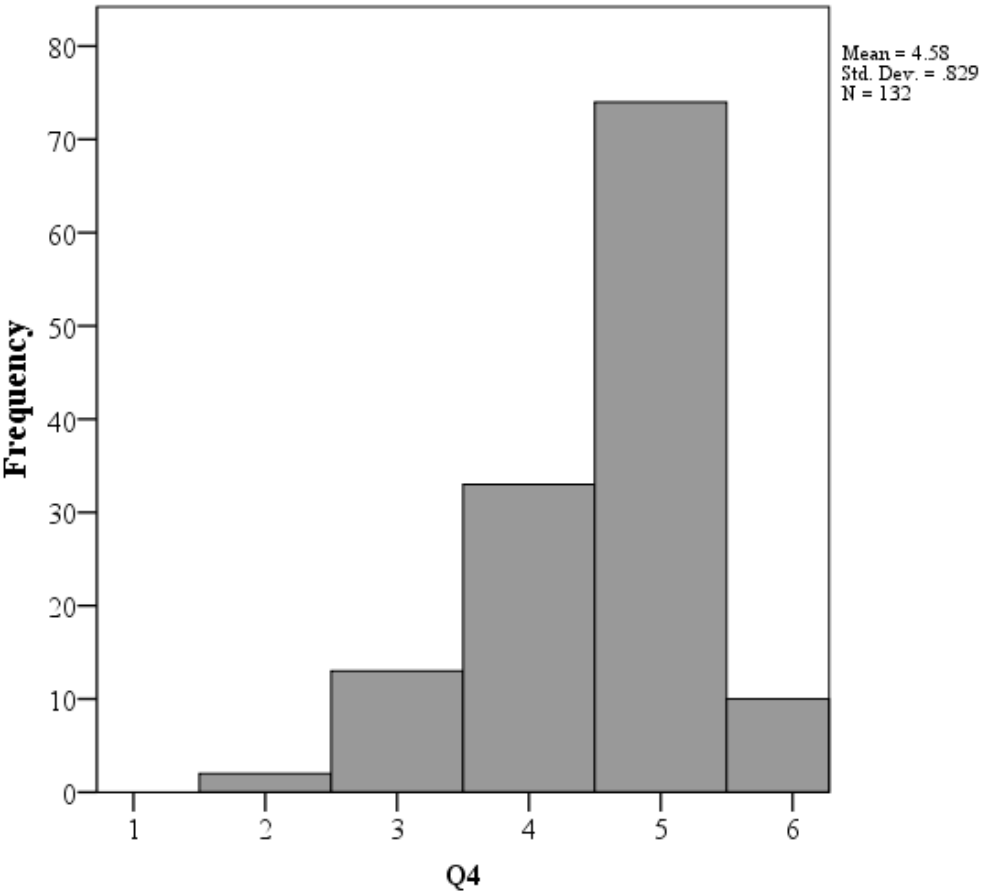
Residents:

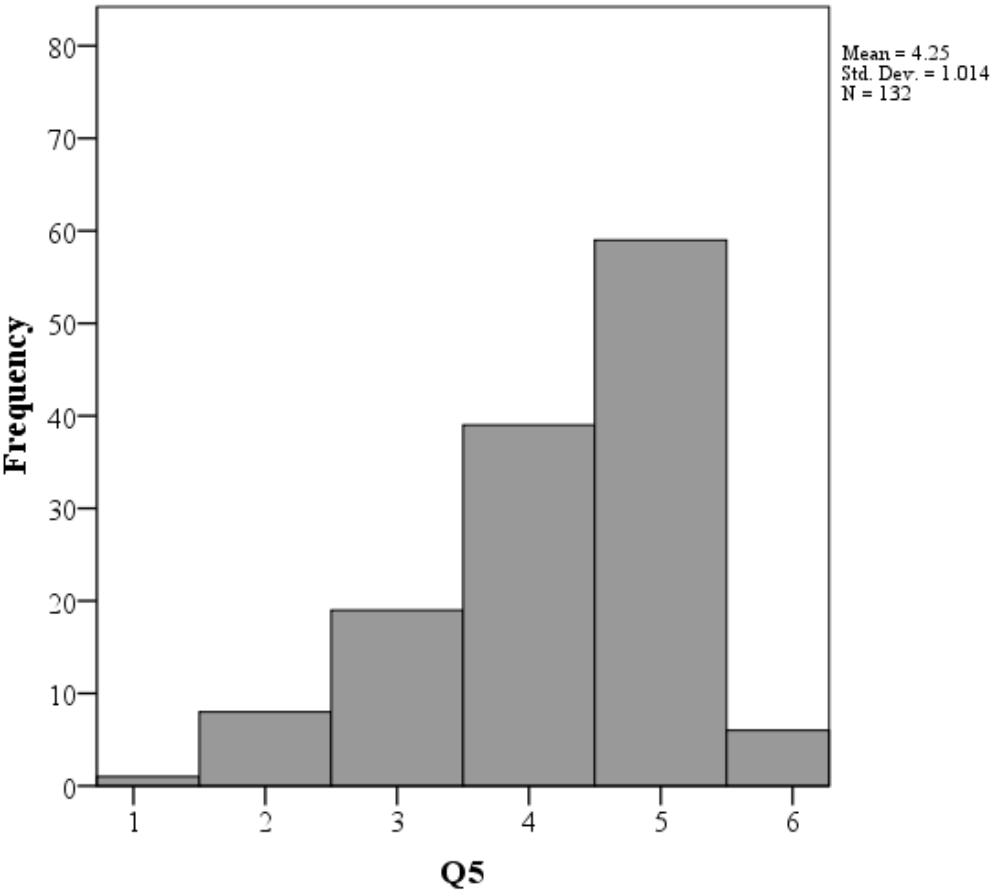
- 1) Did having a staff complete an OCAT on your performance have any impact on the verbal feedback you received in clinic?
- 2) How often did staff review the OCAT with you? How did that occur?
- 3) How did you feel when you received a score of 3 or lower on an item?
- 4) Any other positives of using the OCAT from your perspective?
- 5) Any other negatives of using the OCAT from your perspective?
- 6) Did you find the OCAT easy to use?
- 7) Was the OCAT self-explanatory?
- 8) Would you prefer to have the OCAT on paper or electronically on your iPad? Why?

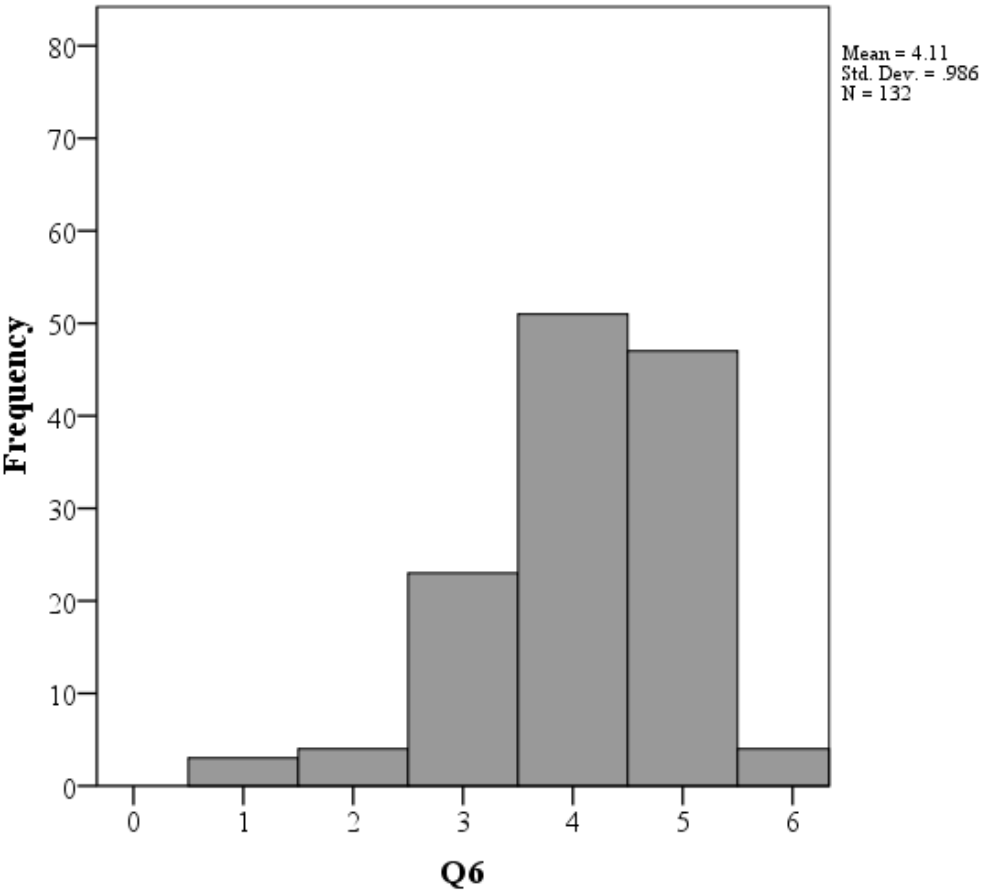
Appendix R: Distribution of Rating Frequencies by Item (1-13)

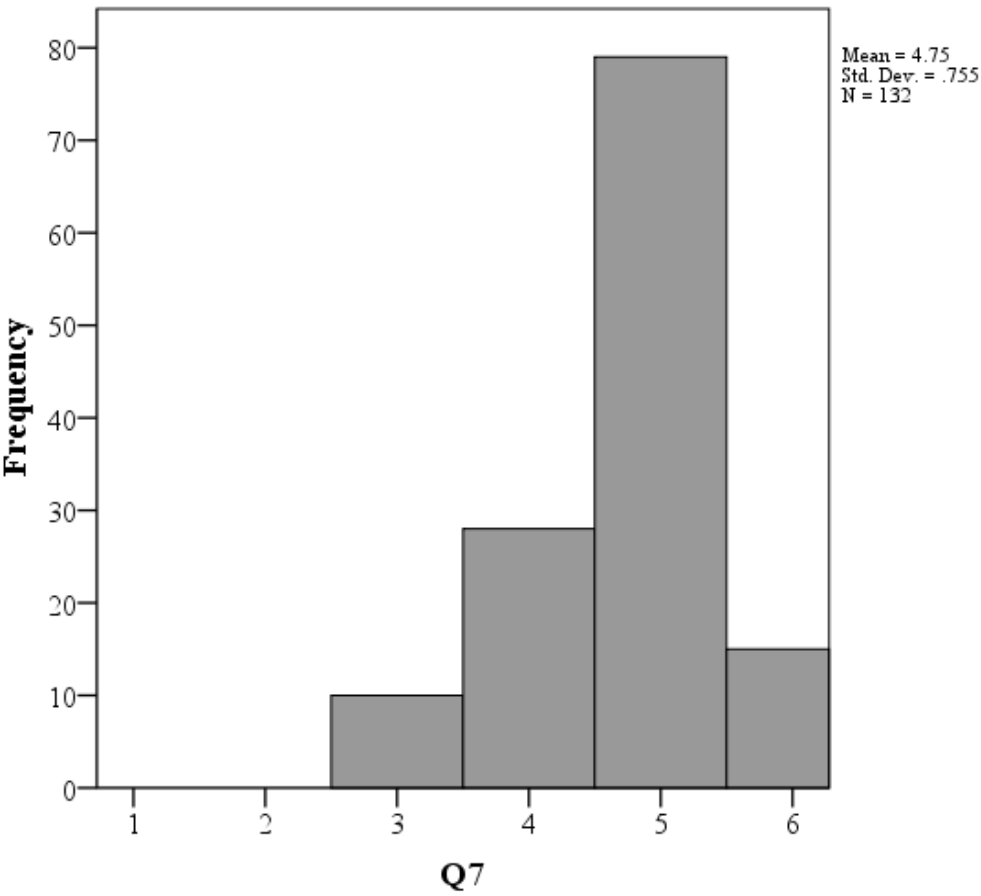


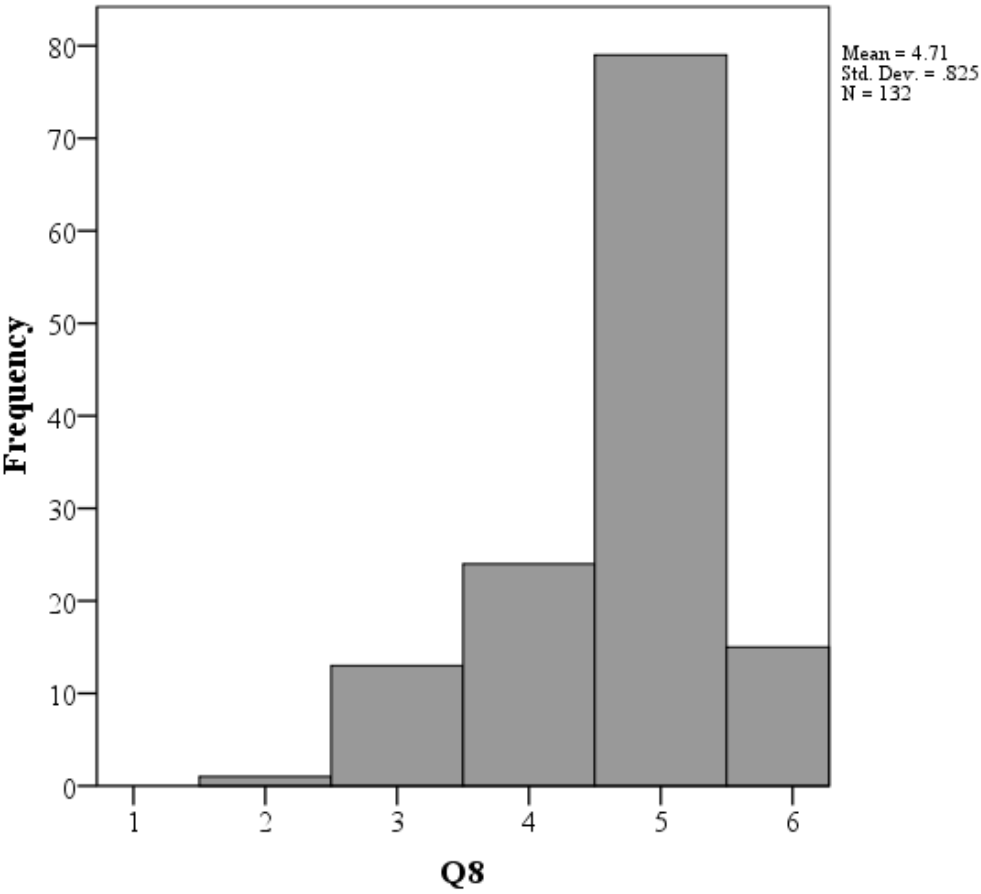


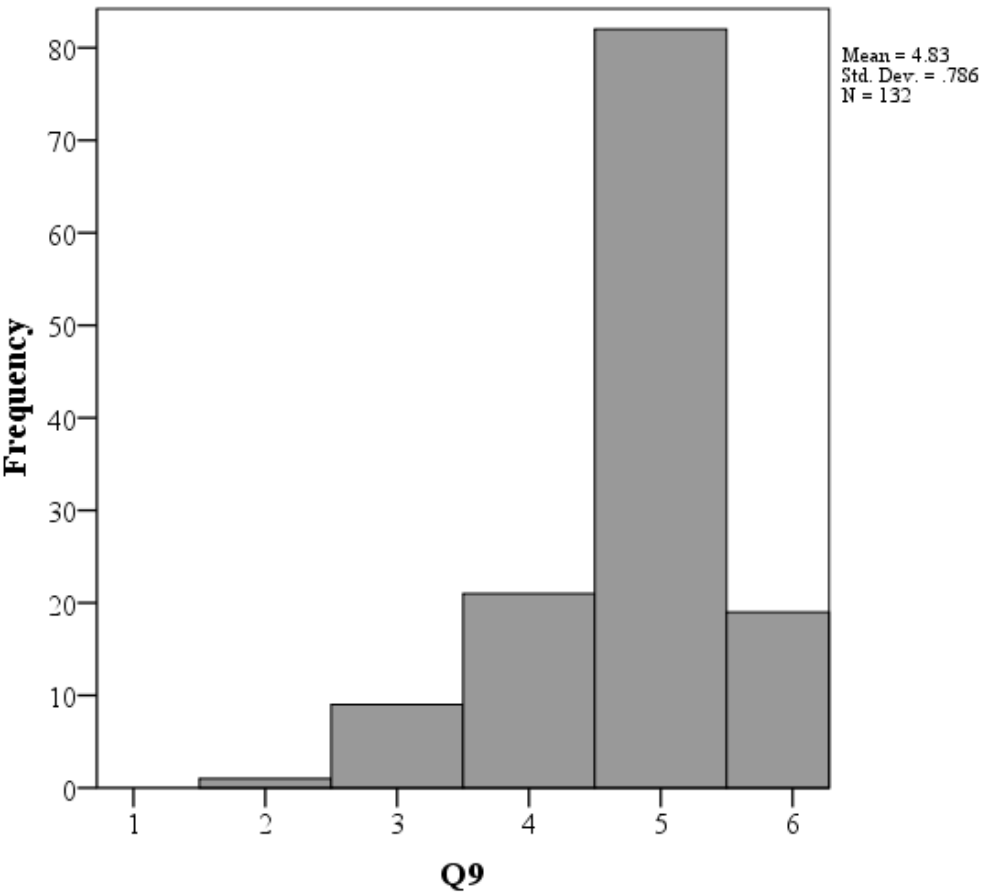


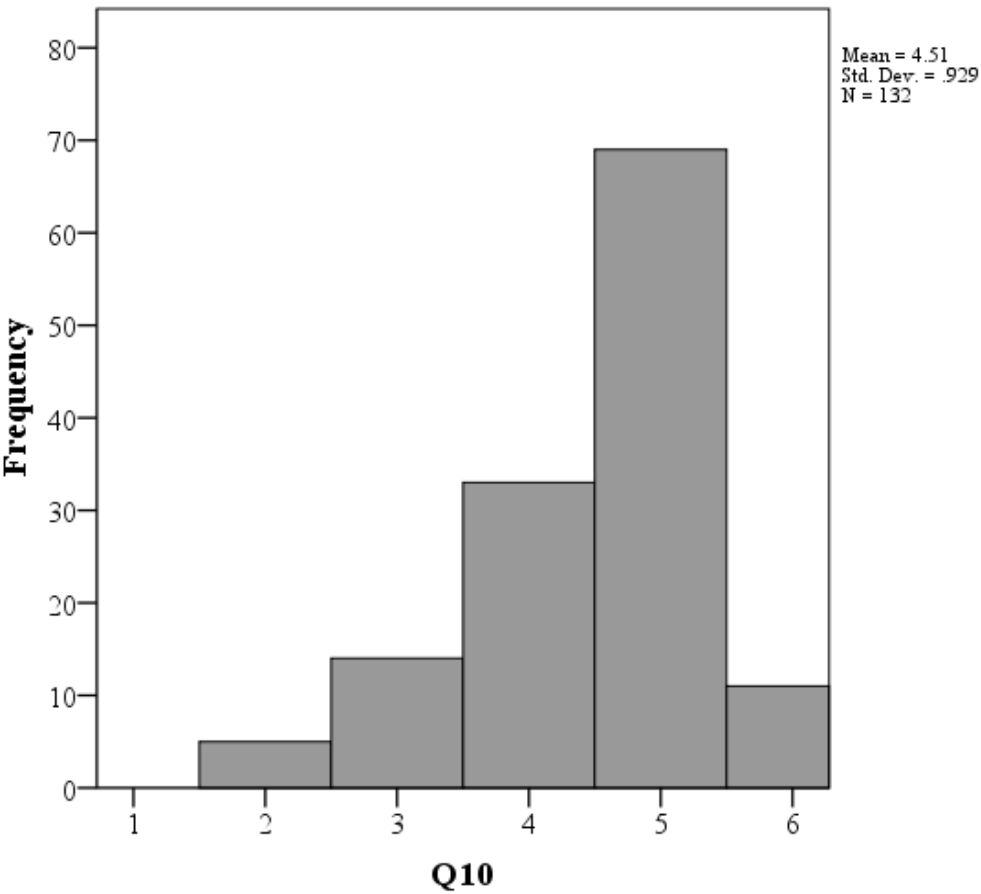


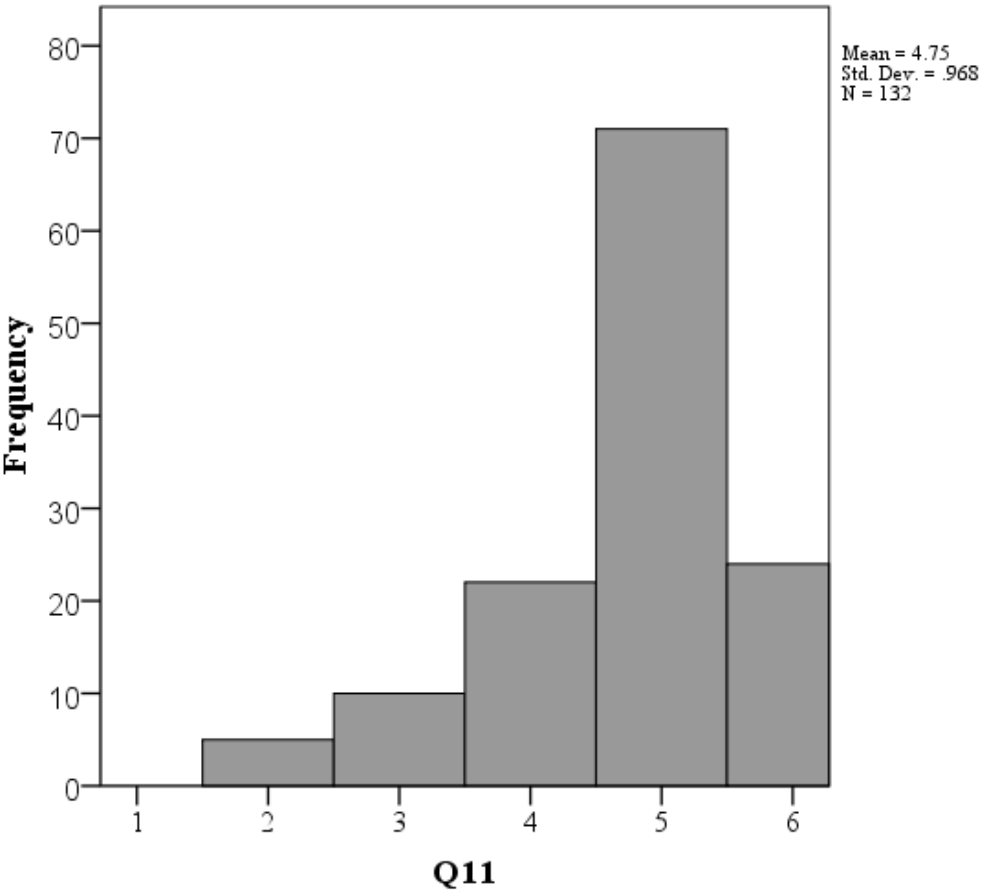


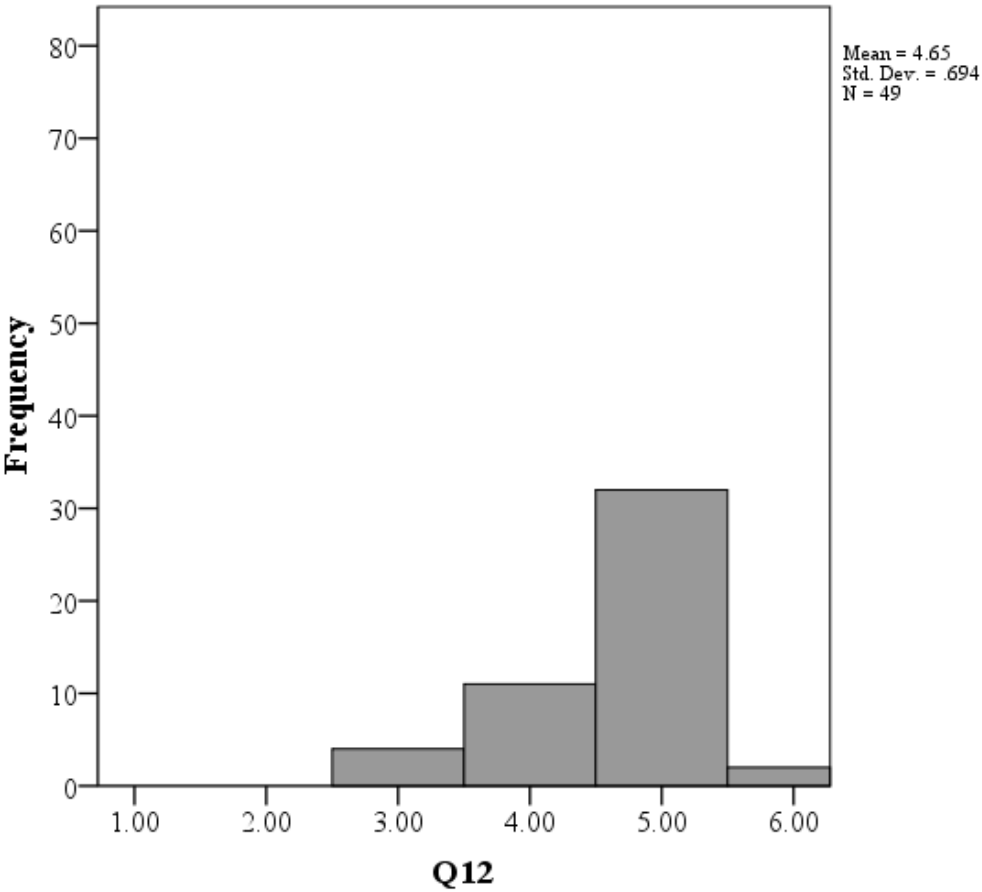


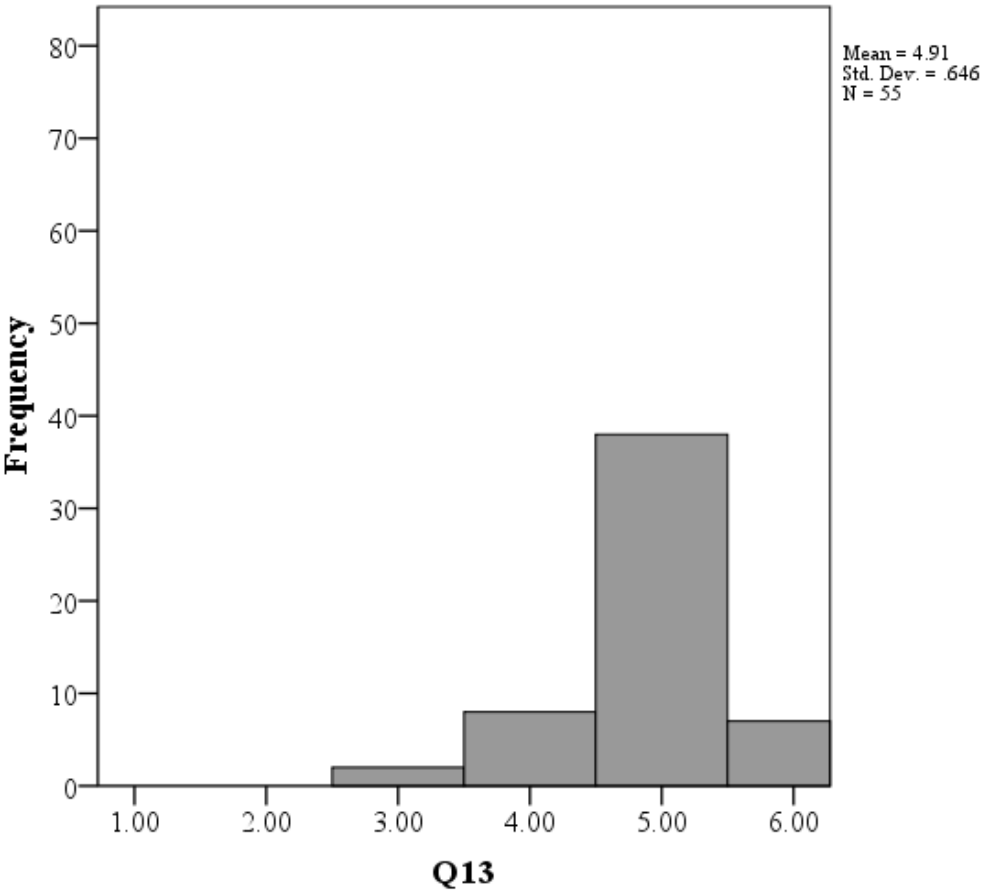












Appendix S: Summary of Future Directions

Short Term	
Shorten the OCAT form to 1 page	Remove Item 1 (clinic preparedness is represented in the other items) Remove Item 11 (attitude) Change Item 11 to a Yes/No question reflecting professionalism displayed in that clinic
Modify item 10	Change “time management” to more clearly represent time management of an <i>entire</i> clinic
6-point rating scale	Change to a 5-point scale on future OCAT
Global assessment item	Reverse orientation of Yes/No answer Wording: re-evaluate the word ‘manage’ as it was found to be confusing
Technical skills subscale	Test whether these items should be - included in future OCAT forms - analyzed as a sub-scale Why was the lowest value assigned a ‘3’. Are staff raters using the OCAT in all clinics involving skills, even if the resident only observes?
Junior resident participants	Need to increase the number of junior residents included in project to properly assess whether the OCAT is an appropriate assessment tool for them as well.
How does the OCAT reflect CanMEDS concepts?	Map the OCAT items and descriptors back to the CanMEDS 2015 roles
Long Term	
Rater training vs. rater orientation	How much rater training is necessary? At least frame of reference training is necessary in future projects to orient staff to how to use the entrustability scale (get away from thinking in terms of PGY level to competence level) and how to think in terms of generalist vs. speciality level (OCAT intended to be used as if training a generalist)
Paper vs. electronic	Will it be used outside of a study if it’s on paper? Especially if there’s no \$10 gift card available for the residents. A system needs to be created to allow for an electronic system.
How do residents accept constructive criticism given on the entrustability scale?	Is it different than a traditional norm-referenced scale?