12-14-2022 5:00 PM

Development of an On-Call Assessment Tool for Competency-Based Surgical Training

Eric C. Mitchell, *The University of Western Ontario*

Supervisor: Grant, Aaron, *The University of Western Ontario*
Co-Supervisor: Ott, Michael, *The University of Western Ontario*
Co-Supervisor: Ross, Douglas, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Surgery

Abstract

Introduction: A central tenet of competency-based medical education is formative assessment of trainees. There are no assessments examining resident competence on-call, despite this being a significant component of resident training and characterized by less supervision compared to daytime.

Methods: A national survey was conducted to evaluate the state of assessment in Canadian Plastic and Reconstructive Surgery programs. An on-call assessment tool was developed based on a consensus group and was piloted over six months. Validity of the tool was examined through qualitative and quantitative methods.

Results: There were 63 tools completed across ten residents and seven staff physicians. Tool reliability was 0.67 and scores were significantly correlated to year of training. Staff and residents considered the tool useful, feasible and acceptable.

Conclusions: The on-call assessment tool has multiple sources of validity evidence to support its purpose of assessing surgical resident competence on-call. Further research is required to assess tool generalizability.

Keywords

Competency-based medical education, assessment, validity, psychometrics

Summary for Lay Audience

The training of a surgeon is complex. In Canada, resident physicians become independent practicing surgeons by being immersed in a five-year work-based curriculum during which they receive didactic teaching as well as supervised hands-on experience. Assessment of knowledge and performance is important in determining whether a trainee is progressing as expected and to provide feedback to enhance future performance. The current residency curriculum focuses on frequent, low-stakes assessment of trainees in the workplace. However, currently there are few assessments during the on-call period when supervisors are often not present, and residents function with greater autonomy.

We surveyed residents and program directors from all Canadian Plastic & Reconstructive Surgery training programs and confirmed that there is a lack of assessment on-call. Residents and program directors believed a more formal way of assessment would be beneficial.

A tool was developed with input from surgeons experienced in medical education that could be used to assess resident performance on-call. This tool was piloted in the Division of Plastic & Reconstructive Surgery in London, Ontario. Ten residents were assessed by seven staff physicians across 63 instances. The tool was able to differentiate between residents of advancing training level. More occasions of scoring will be needed to improve reliability of the tool.

We interviewed four residents and three staff physicians who participated in the pilot to better understand the utility and impact of the tool. Analysis of the interview transcripts revealed there was a positive impact on the amount of feedback given as well as standardization of the feedback process. In addition, the pilot results suggested potential refinements that could be made to improve the practicality of the assessment. However, there was general agreement the tool design was acceptable and useful.

Overall, this thesis offers a better understanding of the landscape of feedback and assessment on-call for surgical trainees. It also provides an assessment tool that can be used to facilitate feedback and learning on-call. Further work should be done to see if this tool is more broadly applicable.

Co-Authorship Statement

MSc Candidate – Eric Mitchell, MD. Resident, Division of Plastic and Reconstructive Surgery, Department of Surgery, Schulich School of Medicine & Dentistry

- Eric Mitchell is the primary author of this body of work. He was responsible for the background research, methodology, facilitating the consensus group, data collection, and analysis of results.

Aaron Grant, MD, MHPE. Associate Professor, Division of Plastic and Reconstructive Surgery, Department of Surgery, Schulich School of Medicine & Dentistry

- Aaron Grant is the supervisor and senior author for this body of work. He was responsible for helping conceptualize the original research idea. He provided guidance on project design, methodology, helped with data collection and analysis, as well as thesis preparation.

Michael Ott, MD, MSc, MHPE. Professor, Division of General Surgery, Department of Surgery and Surgical Oncology, Schulich School of Medicine & Dentistry

- Michael Ott provided guidance on project design, methodology, data analysis, and thesis preparation.

Douglas Ross, MD, MEd. Professor, Division of Plastic and Reconstructive Surgery, Department of Surgery, Schulich School of Medicine & Dentistry

- Douglas Ross was responsible for helping conceptualize the original research idea. He provided guidance on project design, methodology, data analysis, and thesis preparation.

Geoff Norman, PhD. Professor Emeritus, Health Research Methods, Evidence, and Impact, Faculty of Health Sciences, McMaster University

- Geoff Norman provided guidance on methodology and aided in data analysis.

Stacy Fan, MD, MSc. Resident, Division of Plastic and Reconstructive Surgery, Department of Surgery, Schulich School of Medicine & Dentistry

- Stacy Fan was responsible for helping conceptualize the original research idea and designing an initial version of an on-call assessment tool.

Acknowledgments

I would sincerely like to thank all my co-authors who helped guide me throughout this process.

Thank you to Dr Grant for being so instrumental to this whole project across all aspects. You are truly a role model surgeon and teacher.

Drs Ott and Ross provided invaluable expertise and insight as academic surgeons dedicated to improving medical education.

Thank you to my colleagues in the Division of Plastic and Reconstructive Surgery who were so generous with their time and engagement in this project and for supporting me as a co-resident.

Thank you to Kelly for encouraging me to pursue my goals in this thesis, at work and in life.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

CBME – Competency-based medical education

RCPSC - The Royal College of Physicians and Surgeons of Canada

CBD – Competency by design

EPA – Entrustable Professional Activities

ACGME - The Accreditation Council for Graduate Medical Education

OSCRE - Objective structured clinical exams

OSATS - Objective Structured Assessment Tool Skills

O-SCORE - Ottawa Surgical Competency Operating Room Evaluation

GOALS - Global Operative Assessment of Laparoscopic Skills

GRS - Global Rating Scales

GEARS - Global Evaluative Assessment of Robotic Skills

NOTSS - Non-Technical Skills Assessment in Surgery

OTAS - Observational Teamwork Assessment in Surgery

OCAT - Ottawa Clinic Assessment Tool

SLI - Surgeons' Leadership Inventory

WBA – Workplace-based assessment

NGT – Nominal group technique

CTT – Classical test theory

PRS – Plastic and Reconstructive Surgery

PGY – Post graduate year

CanMEDs - Canadian Medical Education Directives for Specialists

PD – Program director

CAMEO – Clinic Assessment and Management Examination – Outpatient

CCAC – Community care assessment centre

ICU – Intensive care unit

TA – Thematic analysis

ITER – In-training evaluation report

Preface

Problem Statement

Competency-based medical education (CBME) is the current standard of postgraduate, specialty medical training in Canada. Success of CBME depends on regular assessment of resident competence across all care settings. Residents spend a significant portion of time throughout training on-call. During the on-call period, residents often have increased autonomy compared to daytime hours and practice decision-making and technical skills without direct supervision. Assessment of on-call performance would be very beneficial to the learning process. However, there are currently no formative assessment tools available to surgical educators to provide feedback to residents based on their on-call performance.

Thesis Objectives

1. To understand the baseline level of feedback and assessment on-call that exists in Plastic and Reconstructive Surgery programs across Canada through a national survey.
2. To identify key elements of surgical resident competence on-call using consensus group methodology.
3. To develop a formative assessment tool to evaluate surgical resident competence on-call.
4. To collect validity evidence for use of the assessment tool within the Division of Plastic and Reconstructive Surgery at our institution.
5. To understand the impact of tool implementation on residents and staff through qualitative interviews.

Chapter 1

1 Literature Review

1.1 A Review of Competency-Based Medical Education and Assessment.

1.1.1 A Background on Medical Training and CBME

The first “modern” surgical residency training program was established by William Stewart Halsted in 1889 at Johns Hopkins University[1]. This was a time-based apprenticeship model with successful completion subjectively determined by Halsted after an average of eight years of training. This program produced true “general” surgeons, prior to the advent of surgical specialization[1]. Over the ensuing century, postgraduate medical and surgical education evolved into a more objective and structured process but fundamentally remained a time-based, apprenticeship model in which trainees spend designated amounts of time obtaining clinical exposure to a specific field. Successful completion has been, and still is, determined by a high-stakes, final summative examination[2]. Recently, there has been a call to improve upon the time-based model to ensure trainees receive and document the required clinical experiences and feedback necessary to become competent in their specialty[3–5]. As it became evident that the time-based training model did not necessarily cover all important skills and experiences, competency-based medical education (CBME) was developed[6].

Competency-based models were first introduced to the broader medical field in 1978 as part of a report to the World Health Organization[7]. The models were described as “an outcomes-based approach to the design, implementation, assessment and evaluation of a medical education program using an organizing framework of competencies"[7]. While proposed to the medical field over forty years ago, only recently has the transition gained momentum and seen increased adoption.

As mentioned, CBME was partially born out of criticisms of previous curricula, in that they failed to ensure all graduates displayed competence in the areas necessary for

independent practice[6]. Other forces behind this transition, include duty hour restrictions and a greater focus on reducing medical errors and enhancing patient safety[8]. Compared to prior training models, trainee advancement in CBME is based on demonstration of competence for specific tasks, as well as knowledge application in real clinical settings. An outcomes-based approach to education such as this is thought to help ensure preparation for independent practice in our era of greater accountability and scrutiny[9]. CBME also promotes learner-centeredness, where trainees are more responsible for their progress and theoretically have flexibility to adjust time dedicated to various clinical duties and tasks[9]. Finally, this model emphasizes frequent formative assessments which is a pedagogical strategy thought to improve learning experiences[2].

1.1.2 Outcomes in CBME

As CBME prioritizes outcomes, a central challenge is how to decide and design relevant outcomes for each specialty. Once these are established, the subsequent challenge is deciding how to assess and evaluate these outcomes demonstrating trainee competence [8–10]. Several organizations have conceptualized what a CBME curriculum may look like. In Canada, The Royal College of Physicians and Surgeons of Canada (RCPSC) has designated seven domains or CanMEDs roles that are considered general competencies for all physicians. These include medical expert, advocate, leader, scholar, communicator, collaborator and professional[11]. In 2015, the RCPSC introduced their version of CBME to Canadian residency programs called “Competence by Design” (CBD) which is a hybrid model of CBME and time-based learning. Important terms and concepts in the CBD model include “competency”, “milestone”, and “entrustable professional activities”. Competency is “an observable ability of a health care professional that develops through stages of expertise from novice to master clinician”, while a milestone is defined as “the expected ability of a health care professional at a stage of expertise”[1]. Entrustable professional activities (EPAs) were originally defined by Ten Cate and are "a key task of a discipline that can be entrusted to an individual who possesses the appropriate level of competence”[1,2]. EPAs are designed as outcome measures specific to each individual medical or surgical specialty[1].

In the United States, The Accreditation Council for Graduate Medical Education (ACGME), has a similar system. ACGME released their six Core Competencies in 1999: patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice[14]. ACGME subsequently updated their curriculum in 2020 to "provide narrative descriptors of the Competencies and sub-competencies along a developmental continuum…"[14]. Many parallels can be drawn between the initiatives of both the RCPSC and ACGME suggesting that CBME is the future of medical and surgical training.

1.1.3 Assessment in CBME

Medical educators need ways to monitor and document trainee progression, provide feedback, and evaluate competency. In this way, outcomes and assessment are intertwined in the CBME model. Assessment in basic terms involves testing, collecting measures of performance which is then utilized to provide feedback[14(p20)]. The toolbox of assessments available to graduate programs for these purposes is diverse and many different outcomes may be assessed. As a trainee progresses in their training, assessments should analyze different and more complex outcomes. One common way of conceptualizing this is through Miller's pyramid of assessment[1]. At the basic level, trainees are assessed if they "know", the next assessment would be if they "know how", then if they can "show how" and finally if they can "do" a defined task[1]. Individual assessment techniques can be thought of as applicable to one or more of these levels. A typical multi-choice question exam is an example of assessing the "knows" level, and written essays are an example of assessing "knows how". These assessment types can be used to evaluate knowledge acquisition and application in theory. Many of the high-stakes and summative assessments that regulatory bodies rely on also occupy these levels. Simulation and objective structured clinical exams (OSCEs) are examples of assessments in the "shows how" levels.

The fourth level of Miller's pyramid, "does", can be more complex to assess. Methods of assessment in the "does" level include chart or electronic medical record review, direct observation in clinical environments, end-of-rotation evaluations, multi-source feedback and case logs[1]. Many medical educators believe CBME assessments should fall primarily

within this level[16,17]. The reasoning for this is twofold: (1) assessment at the "does" level is thought to provide deeper meaning for a trainee and helps build upon the cognitive processes of clinical decision-making[18] which is in line with another CBME principle of prioritizing the use of assessment *for* learning instead of assessment *of* learning[16,19], and (2) these assessments may demonstrate competence for a specific task.

Many of the assessment methods in the "does" level fall into the general category of workplace-based assessments (WBAs). WBAs are designed to assess outcomes within the workplace context, documenting real-world performance, as a proxy for what will be done in independent practice[20]. WBAs are considered a cornerstone of CBME[21,22].

1.1.4 Assessment in Surgery

In 2009, the Division of Orthopedic Surgery at the University of Toronto became the first surgical training program in the world to initiate a competency-based program[23]. Since then, CBME has been increasingly implemented in surgical specialties across Canada and with the adoption of CBD by the Royal College, all surgical programs will transition to this model. The RCPSC has tasked each surgical specialty with defining surgical competencies for their field and design a CBD curriculum based upon these. Surgical competence can be broadly defined as "the ability to apply professional knowledge, skills, and attitude to new and familiar tasks in all clinical environments"[24]. Both technical and non-technical skills must be learned and applied. For surgical specialties, procedural skills in the operating room are the most obvious, but care of inpatients, seeing patients in clinic, and in the emergency department are also important. Outcomes and assessment in surgery must cover technical and non-technical skills ideally in all clinical environments.

There are a plethora of assessment tools specific to surgery as no single tool can assess all dimensions of competency and each tool comes with its own benefits and limitations[9,25]. A systematic review of technical skills assessment tools in surgery found the most commonly used tool was the Objective Structured Assessment Tool Skills (OSATS), which is used to assess technical skills through simulation and has been applied across numerous surgical specialties[26,27]. Other tools to assess technical skills include the Ottawa

Surgical Competency Operating Room Evaluation (O-SCORE)[28], the Global Operative Assessment of Laparoscopic Skills (GOALS)[29], Global Rating Scales (GRS)[30], and the Global Evaluative Assessment of Robotic Skills (GEARS)[31]. Tools examining non-technical skills in surgery include the Non-Technical Skills Assessment in Surgery (NOTSS)[32], the Observational Teamwork Assessment in Surgery (OTAS)[33], the Ottawa Clinic Assessment Tool (OCAT)[34], the Surgeons' Leadership Inventory (SLI)[35], among others. As is evident, surgical educators have a varied toolbox of assessments to choose from and identifying the appropriate tool for each task is challenging. Having a good understanding of the principles of assessment can help.

1.1.5 Principles of Good Assessment

Research in the medical education literature suggests several ways to design and optimize assessment tools. Norcini et al. 2011 outlined seven consensus criteria for good assessment: 1) validity or coherence, 2) reproducibility or consistency, 3) equivalence, 4) feasibility, 5) educational effect, 6) catalytic effect and 7) acceptability[14]. The authors acknowledged that no single set of criteria applies equally well across all situations and the weight placed on each criterion will differ depending on the assessment and stakeholders. For example, with a high-stakes licensing examination, validity and reliability are more important; conversely, for certain formative assessments, the educational effect or feasibility may be prioritized. The idea that the weight associated with each criterion differs depending on the situation is similar to the Utility of Assessment Methods model outlined by van der Vleuten in 1996[36]. In that paper, the five contributing variables to utility were reliability, validity, educational impact, acceptability, and cost. Within this model, an individual assessment tool could be thought of as having different weights assigned to each variable, which contribute to the overall utility of the tool.

Lockyer at al. discussed core principles of assessment specific to CBME[16]. These included using assessment *for* learning, active engagement of learners and designing assessment within the "does" level of the Miller pyramid. Additionally, they highlighted the importance of using multiple methods of assessments, multiple assessors who are appropriately selected and trained, and employing psychometrics. Multiple methods of

assessment are important to compensate for the shortcomings of any one technique and both qualitative and quantitative data have their role in CBME assessment. Multiple assessors are needed to mitigate assessor bias, leniency, and halo effects. The discussion surrounding psychometrics included the changing role and definitions of aspects like validity and reliability[16].

1.1.6 Assessment Challenges in CBME

While principles exist to aid in good assessment design, many challenges still remain in assessment implementation in CBME including time constraints and feasibility[14,37–40], as well as a lack of understanding of purpose and underlying doubts about their ultimate educational value[37,41]. Forty-one percent of surgical trainees and surgeons in a United Kingdom survey found the time required to complete mandatory WBAs online negatively impacted training overall, while only 6% believed it positively impacted training[38]. A narrative literature review of articles from 2005 to 2015 found consistent trainee concerns regarding the time required to complete WBAs[37]. Many WBAs are considered too bureaucratic, complex and too much of an administrative burden[39]. These issues are only amplified in CBME with its emphasis on more frequent assessments. There is a fine line between too few assessments and assessment overload[42]. To help avoid assessor fatigue, it is recommended assessments should be available for the right purpose at the right time through optimal use of multiple assessors and tools[16].

Another barrier to WBA implementation has been a lack of misunderstanding of the purpose of WBA. The high-frequency and low-stakes assessments in CBME are meant to be formative. While designed for learning and feedback, trainees may still think of them as summative in nature[39,43–45]. As a result of this, trainee engagement is negatively affected, and trainees may avoid discussing cases that are more complex or difficult[37]. Encounters which should be helping learners refine their skills and improve their knowledge are avoided, interfering with the intended educational effect[41].

1.1.7 The On-Call Period

Progressive independence is a pillar of clinical training and medical education[46]. It is the process of increasing trainee independence in patient care delivery while simultaneously

decreasing levels of supervision[46]. Traditionally, the highest levels of independence and autonomy experienced by residents has been during the on-call period, when supervising physicians are often not in hospital. Residents on-call often function semi-autonomously. Supervising physicians should be easily reachable when needed, but even so, the act of residents requesting staff support is complex, highly context dependent and depends on trainee and supervisor characteristics meaning residents may manage more than they would compared to if staff were present[47,48]. Although lacking robust empirical evidence, perceived benefits of trainee independence include an association with themes of increased trainee confidence, readiness for independent practice, and the development of clinical decision-making skills and professional identity . A decrease in trainee autonomy could have the unintended consequences of producing clinicians with limited experience functioning independently. Our knowledge on the true impact of resident autonomy on-call is limited[1,2], but anyone who has been through the process of seeing a patient overnight without a supervising physician present in hospital for immediate backup can relate to how impactful an experience it can be.

To better understand the impact of the on-call period, we must understand what residents "do" overnight and on-call. A time-motion study looking at how general surgery residents at the University of North Carolina at Chapel Hill teaching hospitals spend their time on-call found 20% of the night was spent evaluating patients, 57% on activities of daily living and the rest of the night was split between communication, pages, procedures and other miscellaneous items[52]. Residents in this study completed an average of eight patient evaluations per night. A 2014 study examining the experience of plastic surgery residents and fellows on-call found that most received 6-9 calls per night, and a large majority (83.6%) reported they "mostly" or "always" were called back into the hospital after leaving[53]. Most programs in this study used a "home-call" set-up, considered by some to reflect an attending surgeon's practice more realistically, preparing residents for that aspect of independent care[53]. A time-motion study of general surgery residents at a Canadian academic centre found that in a 15-hour call period (from 17:00 to 08:00) there were, on average, 2 hours of direct care and over 6 hours of indirect care, which included medical record use, documentation, handover, and team communication[54]. Another group examined the activities performed by residents on surgery "night-float" at the University

of Oklahoma Health Sciences Center[55]. It found most of the time was spent doing educational activities and residents expressed overwhelming support in favour of the night float system as an educational experience and a way to benefit from level-appropriate autonomy[55]. In Ontario, residents can spend up to one-in-three nights on-call as per Ontario Provincial contracts, meaning the on-call period encompasses a significant portion of the educational experience. The overall on-call period is a unique opportunity for resident learning under conditions of increased autonomy. To optimize the educational impact of autonomous practice, including that which is done on-call, instruments are necessary to provide meaningful feedback for trainees[50]. However, there are, no widely used assessment or feedback tools to look at on-call performance or competence.

A group of Internal Medicine educators proposed the use of a 360-degree assessment tool as a way to assess resident performance overnight when not being directly observed, but this has not been trialed yet[56].

The University of Cincinnati and University of Iowa Ophthalmology programs developed the On-Call Assessment Tool (OCAT) to evaluate, what the authors described as, three critical aspects of on-call performance: patient care, timeliness, and sense of urgency[57]. These aspects were identified based on a literature review. The study examined the face, content, and discriminative validity of the tool, but did not provide evidence based on modern validity theories and did not examine reliability. This is the only tool identified as being specific to the on-call period; however, it is not widely applicable to other surgical specialties or hospitals and was not developed using modern medical education assessment development principles.

1.1.8 Direct and Indirect Supervision

An obstacle to meaningful assessment of residents based on their on-call performance is the lack of direct supervision. In general, medical educators contend that assessment of resident clinical activities should ideally be done after direct observation[18,58,59]. Trainees and physician supervisors also agree on the importance of direct observation[60]. The definition of direct observation varies, but one group defined it as "the active process of

watching learners perform in order to develop an understanding of how they apply their knowledge and skills to clinical practice"[59]. Direct observation, as opposed to feedback or assessment based on indirect observation or inferences, is thought to improve the reliability and validity of clinical performance ratings and assessments[61], and pedagogically it's use makes sense.

Arranging for direct observation in the clinical setting, even during regular daytime hours, can be difficult. Feedback based on direct observation may not be feasible given workflow demands, a desire for increasing trainee independence, or because many tasks are inherently not amenable to direct observation[22,23]. Despite knowing that it is important and valued, direct observation in the workplace happens infrequently[62–65]. Direct observation also comes with its own drawbacks as some trainees express experiencing significant anxiety and discomfort associated when being observed[42,59]. Given these concerns, educators recognize that indirect observations can and should play a role in assessment and provision of feedback alongside direct observation[22,23]. More importantly, the most unique feature of on-call performance is the lack of direct supervision with resultant learner autonomy.

1.1.9 Summary

CBME is the standard of postgraduate training in Canada. CBME emphasizes assessment of competence and knowledge application in real clinical settings as well as the application of frequent low-stakes formative assessments that occupy the "does" level of Miller's pyramid of assessment. WBAs have been designed to meet these criteria, evaluate outcomes in the workplace context and document performance as a proxy for what will be done in independent practice. WBAs are considered the cornerstone of CBME. The pressure and burden on programs to regularly evaluate their trainees is high within CBME. As a result, it is important to ensure the assessments being used are of high quality (in terms of validity, feasibility, educational impact, etc.), and that they actively engage learners. There should also be multiple methods of assessments and multiple assessors should be used to reduce the burden on any one individual.

An aspect of clinical training that historically has had limited assessment or evaluation is the time residents spend on-call. The on-call period encompasses a significant portion of workhours across the course of a residency and is characterized by indirect supervision and increased autonomy compared to daytime hours. Despite this, there are no assessments designed to assess performance on-call. Assessments can be used to provide meaningful feedback to optimize the educational impact of semi-autonomous clinical practice on-call. This project aims to build on research in the field of resident assessment specific to the on-call period.

1.2 Review of Methodology

This section examines consensus group methodology and psychometric analysis of assessment tools in further detail.

1.2.1 Consensus Group Methodology

Empirical evidence in the field of medical education is often limited[66]. In fields where published literature is incomplete or inadequate, consensus group methods provide a means of harnessing and synthesizing the insights of experts[66,67]. Consensus group methods like Delphi and Nominal Group Technique (NGT) are widely used in the field of medicine and medical education[66,67]. The rationale for using consensus groups is based on several assumptions about group decision making compared to individual decision making[68]. These include: a selected group of individuals is likely to provide some level of authority, decisions are improved when they undergo group challenges and when members justify their views, and the likelihood of making a wrong decision is lower with more rather than fewer people[68]. Structured methodology is needed to impart credibility, but also to organize potentially complex and varied opinions and ensure each group member can contribute meaningfully. Ultimately, if the group arrives at a consensus, the consensus can be accepted and applied by others going forward. Some key features of formal consensus groups as outlined by Humphrey-Murto et al. are: anonymity, iteration, controlled feedback, statistical group response and structured interaction[69]. Downsides of consensus group methods include the potential for bias from selection of participants, or results may end up capturing collective ignorance if the group is not appropriate[68]. Additionally, consensus groups cannot be used in place of rigorous empirical evidence, but rather should be thought of as a first step in the process of further data collection and comparison against actual observable events[67,70].

1.2.1.1 Delphi Method

Within medical education literature, 75% of papers that employ consensus group methods use the Delphi and modified Delphi methods[66]. The Delphi method often involves mailing out a survey or questionnaire to expert participants, with samples ranging from 4 to 3000 participants[71]. After responses are collected by return mail, ratings are combined

and sent back to the participants to review. At that time, they can usually re-rank. The number of iterations or times the survey is sent out to the sample depends on the project. A benefit of the Delphi method is the ability to involve large numbers of participants who are unable to meet in person or amongst whom discussion is not necessary[67].

1.2.1.2 Nominal Group Technique (NGT)

Nominal Group Technique functions as a structured group interaction involving 5 to 12 participants. It is often used for item generation and provides an opportunity for face-to-face discussion. Item generation is done based on a *nominal question* related to the overarching issues or construct of interest and occurs in a round-robin fashion. Item generation continues until no further original ideas are provided. This is followed by group discussion and justification of each item in turn. Usually, members then vote anonymously on the items and the voting results are fed back to the group. Voting may continue for a set number of rounds or until consensus is achieved. NGT has been used to develop assessment tools[35,72], inform curricula[73–75], and establish medical education priorities[76]. A limitation specific to NGT is that it is typically suited to examine only one idea or question in a single session[70]. Table 1 compares the Delphi and NGT methods.

Table 1. Overview of consensus group methods

	Consensus Group Method	
	Delphi	**NGT**
Format of Meeting/Voting	Mailed	In person
Approximate Number of Participants	4-3000	4-12
Private Decisions Elicited	Yes	Yes
Formal Feedback on Group Choices	Yes	Yes
Structured Interaction	Yes	Yes
Group Discussion	No	Yes
Common Uses	Curriculum development	Item generation

Adapted from Jones & Hunter 1995[67], Humphrey-Murto et al. 2017[69]

1.2.2 Validity in Medical Education

Validity is considered by many as the most important characteristic of assessment data, as without it, assessment data has little to no meaning[77]. All assessments require validity evidence, which is used to support or refute an interpretation assigned to assessment results[77]. For example, imagine we are using an assessment tool to help decide about whether a resident is competent to perform a specific operation. The score a resident obtains may suggest they are indeed competent and this score or assignment of

"competence" can be thought of a hypothesis generated based on the tool. To support or refute the hypothesis generated through use of the tool, we require validity evidence. Validity is required because most assessments deal with specific constructs or "intangible collections of abstract concepts or principle"[77,78]. There is no perfect way of scoring or assessing an intangible concept, but validity evidence shows us how close we are to approximating it. Like any other hypothesis, a validity hypothesis can be tested by collecting evidence and organizing it into a validity argument. Using the example of the operating room assessment tool, resident competence is a relatively abstract concept with multiple interpretations but using a "valid" assessment tool designed to assess competence allows us to approximate.

Validity in the classical framework was divided into three components - content validity, criterion validity, and construct validity[79]. The classical framework has been replaced by contemporary theories, best described by Messick[80] and Kane[81]. In these theories, all validity relates to construct validity. Messick defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests"[80]. He described five distinguishable aspects or sources of validity evidence: content evidence, response process evidence, internal structure evidence, relations to other variables evidence, and consequences evidence. An additional sixth aspect, generalizability, was later added and is synonymous with reliability[82]. A description of these aspects follows.

Content evidence: A description of steps taken to ensure that assessment content (questions, prompts, items, instructions, etc.) reflect the construct the assessment is intended to measure[83]. Evidence examples include obtaining expert review, ensuring revisions to content as needed, and basing the assessment on previously used tools [77,84].

Response evidence: An evaluation of how well rater or examinee responses align with the intended construct[83]. Evidence examples include assessor training, use of construct-aligned scales, familiarity of stakeholders with formatting, and providing a rationale for combining scoring methods[77,84].

Internal structure evidence: Examines the relations among the individual items and how they relate to the underlying construct[72,73]. Evidence examples include conducting an item analysis looking at inter-item correlation or item-total correlation, or generalizability studies[77,84].

Relations to other variables evidence: Examines the associations between assessment scores and other measures which are thought to share a specific theoretical relationship[83]. Evidence examples include association with another clinical care measure, educational data or training level[84]. The relationship can be positive (e.g., using two measures which look at the same construct) or negligible (e.g., for two measures which should be independent).

Consequences evidence: Focused on the impact or consequences of an assessment itself and the decisions and action that result[83,85]. Evidence examples include the general impact of scores on students and society and consequences for future learning and teaching and acceptability[77,84,85]. Importantly, the positive consequences of an assessment tool should outweigh the negative consequences, and negative consequences should not result from another source of test invalidity.

Reliability: Refers generally to the consistency of an assessment and can reflect consistency between different raters, items within a tool, stations, different occasions, etc. depending on the type of assessment[78]. It is a necessary but insufficient condition for validity. Derived from Classical Test Theory (CTT), reliability can be defined as the ratio between “true variance” and “total variance”. Reliability can be reflected in many ways, through reliability statistics or coefficients, the most well-known being Cronbach’s alpha, which measures internal consistency. In the case of an assessment tool looking at surgical resident competence, Cronbach’s alpha reflects how strongly each individual item on the tool depends on the surgical resident’s competence. A higher alpha value indicates greater alignment around the construct of interest (competence). An alpha of 0.9 or higher has been suggested as important for very high-stakes tests (e.g., licensing exams), whereas an alpha of 0.7 to 0.79 is acceptable for lower-stakes assessments (e.g., formative assessments delivered locally)[86,87]. However, even tools with scores below 0.70

may still be useful as one component of an overall assessment program. Generalizability Theory (G-Theory) is an extension of CTT and is used when assessments become more complex[78,88]. G-Theory, unlike CTT, can examine more than one source of variance, which is particularly useful for complex assessments like WBAs where score variance can arise from raters, subjects, tool items, different occasions, etc. Each potential source of variance is called a *facet*. Conducting a Generalizability Study (G-study) provides outputs of the variance components by means of a repeated measures ANOVA. Further steps can be taken in a Decision Study (D-study) to look at ways in which theoretically altering the various *facets*, like changing the number of raters, or number of items on a scale, may change reliability.

As mentioned, Kane has contributed a modern framework to study validity as well, writing that validity should be thought of, not as a number, but rather as an argument that supports the final judgement[81,89]. There is a clear focus in Kane's work on creating a purpose statement and accompanying assumptions for an assessment, which provides a "interpretation / use argument" or IUA. The subsequent step is to evaluate the IUA with logic and data to create a "validity argument". The work outlines four categories of assumptions or inferences to consider: scoring, generalization, extrapolation, and decisions.

For the purposes of this thesis report, Messick's framework will be used[80]. There are several other important considerations when looking at validity. One is the notion that validity is on a continuum and is not dichotomous. Within a study it is possible to have validity evidence to support one inference, but not another, and validity is not all or nothing[90]. Gathering validity evidence should be an ongoing process, not something that is carried out initially when implementing or designing a tool and then forgotten. An example of this ongoing process is outlined in a paper by Kinnear et al., in which they examine their process of collecting validity evidence for a WBA[84]. Another consideration is that the quantity or extent of validity evidence required depends highly on the purpose of the assessment. When the attribute being studied is simple and straightforward, the evidence required should be small, but reasonable, while more complex attributes necessitate stronger evidence[78]. For example, a summative assessment like a RCPSC

exam requires more validity evidence than a low-stakes formative feedback assessment tool. The type of evidence required also differs based on the purpose of the tool. While internal structure may be less important for a formative tool, the consequences of the tool may be critical to ensure the tool is having its intended educational impact. Norcini et al. emphasized this point, even proposing that assessment characteristics like catalytic effect, educational effects and to a slightly lesser extent, acceptability and feasibility, are equally as important as validity when looking at formative assessments[15].

Despite the consensus that validity is essential, the validity evidence presented for medical education assessments tools is, in general, scarce[27,91,92]. This holds true across multiple assessment categories including those looking at surgical technical skills[27], simulation-based assessment[91(p20)], and tools used for direct observation of clinical skills[92]. A systematic review of simulation-based assessment in medical education found that out of 217 studies reporting 2 or more sources of validity evidence, 24% of studies made no reference to a validity framework, 69% cited an outdated or limited framework not well accepted, and only 3% referenced Messick's framework, while none referenced Kane's[91].

1.2.3 Summary

Consensus group methods like Delphi and NGT are regularly used in the field of medical education. These techniques harness the insights of a group of experts when published empirical evidence is limited. NGT works a structured group interaction and is often used for item generation based on a nominal question or construct of interest. It is typically suited to examine only one idea or question in a session. Delphi method often involves a larger group of participants who are unable to meet in person and is most often used for curriculum development in medical education.

Validity is an essential component of assessment design and implementation, as without validity assessment data has little meaning. Despite this premise, the validity evidence for most assessment tools used in medical education is scarce. There are multiple frameworks used to describe validity. In this thesis we will apply Messick's framework which focuses on 5 main sources of validity evidence: content, response process, internal

structure, relations to other variables and consequences. We will use consensus group methods and apply the principles of validity theory in the design and testing of our assessment tool.

Chapter 2

2 Assessing the Current State of Feedback and Assessment On-Call in Canadian Plastic and Reconstructive Surgery Programs

2.1 Introduction

There are currently no established tools used for formative assessment of residents on-call. Assessments are a critical component of an outcomes-based competency curriculum and should be designed primarily to facilitate learning and feedback. To improve feedback and learning on-call, it is important to understand what the current state of feedback and assessment is. Much of the current literature focuses on understanding what activities residents do on-call, in particular comparing the proportion of "educational" and "non-educational" activities[52,93,94]. There is some literature on the state of feedback and learning more specifically on-call; however, this is largely based on night float call designs. One study collected survey responses from residents on a surgical night float system and found that nearly half of resident respondents say they receive feedback regarding clinical decisions made at night, and when asked after their call shift, 46% strongly agreed they receive feedback regarding procedural skills on-call[55]. Little is known about the state of feedback and assessment specific to the home-call system which the majority of North American plastic and reconstructive surgery programs employ[53]. There are critical differences among the various on-call shift designs of night float vs. in-house call vs. home-call. Differences include team size, staff presence, general level of supervision, etc. and the differences make it impossible to rely on evidence from other designs to inform the situation for home-call. Drolet at al. reported on some general trainee perceptions regarding the home-call set-up, in the context of work-hour restrictions, but further information is needed[53].

Informal discussion with residents in the Western University program identified a gap in on-call feedback. We wanted to establish whether this gap exists in other programs as well. Therefore, the purpose of this study is to examine the current state of feedback and assessment on-call in Canadian PRS programs. We will investigate whether residents are

interested in receiving more feedback and what areas of practice they want this feedback to encompass. This study will provide guidance for the development of our assessment tool which will be discussed in the next chapter.

2.2 Methods

In order to understand the current state of feedback and assessment on-call in Canadian PRS programs, an online survey was developed to nationally sample residents and program directors. We sought to outline the volume of on-call shifts per month, clinical volume during a shift, typical supervision levels by staff, and resident satisfaction with feedback given based on patient encounters on-call. We wanted to identify any current methods of on-call assessment being used. We also wanted to see whether need for an on-call assessment tool was expressed by other residents or program directors. Questions within the survey were developed based on these goals and review of the literature. We included both closed and open-ended questions.

This survey was approved by the Research Ethics Board of London Health Sciences Centre and Western University (Appendix 1).

Participants: Inclusion criteria included any Canadian Plastic and Reconstructive Surgery (PRS) residency program director or postgraduate year (PGY) 2-5 residents. At the time of data collection, there were 138 potential participants, 126 resident participants and 12 program directors, distributed across 12 residency programs. Resident numbers were confirmed through communication with each program's administrative assistant. Exclusion criteria included being in the PGY-1 training year because they may not have had sufficient on-call exposure for informed survey completion based on when the survey was open for completion (Sep-Oct 2021).

Recruitment: An email containing an invitation to complete the survey was sent to program administrators and then distributed to all potential participants (see Appendix 2 for email script). One reminder email was sent two weeks after the initial invitation.

Data Collection: The anonymous survey was created using REDCap (Research Electronic Data Capture) software. See Appendix 3 for the survey.

Analysis: Data was analyzed using descriptive statistics using Excel® (Microsoft).

2.3 Results

The overall survey response rate was 30% (n = 41/138). Forty-two percent (5/12) of program directors responded and twenty-nine percent of PGY2-5 residents (36/126). One resident participant did not fully complete the survey and therefore was left out of the reported results that follow.

Participant Characteristics and On-Call Burden

The characteristics of resident participants and reported on-call burden are presented in Table 2.

Table 2. Resident survey participant characteristics and on-call burden

	n (%)
PGY-level	
2	14 (39)
3	8 (22)
4	9 (25)
5	5 (14)
Number of on-call shifts per month	
0-1	0 (0)
2-4	4 (11)
5-7	21 (58)
8-10	11 (31)
Number of consults per on-call shift	
0-1	1 (3)
2-4	31 (86)
5-7	4 (11)
8+	0 (0)

Excluding cases that go to the operating room, 86% (30/35) of residents reported they usually received indirect supervision (defined as communication with staff remotely by phone) on-call. One resident reported usually receiving direct supervision and one resident reported no supervision (no communication with staff until morning handover). Three residents reported supervision was usually provided by a senior resident as opposed to staff. Of the residents receiving indirect supervision, 63% (19/30) often discuss management plans with staff, 23% (7/30) always discuss and 13% (4/30) rarely discuss.

The reported level of resident satisfaction with the amount of feedback and quality of feedback given on-call by staff can be seen in Figures 1 and 2, respectively.

Figure 1. Resident satisfaction with amount of feedback

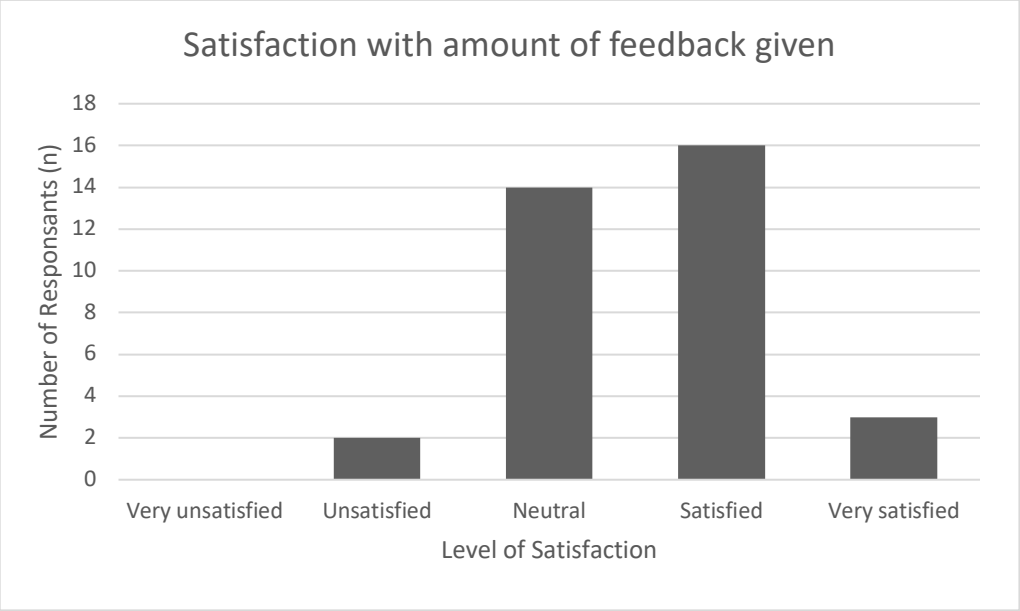


Figure 2. Resident satisfaction with quality of feedback

Eighty-nine percent (31/35) of residents noted their program does not currently use a tool or form to provide feedback to residents on-call. Three residents reported their program uses a tool or form and described these as monthly reviews, quarterly reviews, or EPAs. One resident was unsure.

Sixty-six percent (23/35) of residents believed they would benefit from having a more formal method of receiving feedback, while 17% (6/35) disagreed and 17% (6/35) were unsure.

Program Director Responses

Four program directors agreed there is more room for feedback to be given to residents based on on-call performance, with one program director being unsure. Four program directors said they do not use a tool or form to provide feedback to residents based on on-call performance. The one program director who did use a tool or form used a locally developed tool to provide feedback on procedures. Three program directors believed residents and staff would benefit from having a tool or form available, while two were unsure. Program directors reported which CanMEDs roles they considered to be important when assessing residents on-call (Table 3)[12].

Table 3. Voting results on importance of CanMEDs roles

CanMEDs Role	No. responding Yes (%)
Medical Expert	5 (100)
Communicator	5 (100)
Leader	5 (100)
Collaborator	4 (80)
Professional	4 (80)
Health Advocate	2 (40)
Scholar	1 (20)

Residents were asked which aspects of their on-call encounters they would want feedback on, while program directors were asked which aspects they believe are important to assess and provide feedback on for residents. These results are reported in Table 4.

Table 4. Resident and program director responses to on-call feedback aspects

	Residents	**Program Directors**
Item	No. responding Yes (%)	No. responding Yes (%)
Overall clinical judgement	30 (85)	5 (100)
Clinical outcomes	29 (83)	3 (60)
Development of management plan	25 (71)	5 (100)
Technical-related decisions	23 (66)	2 (40)
Overall patient satisfaction with encounter	11 (31)	2 (40)
Communication with medical team (documentation, handover)	6 (17)	5 (100)
Patient satisfaction with communication	5 (14)	2 (40)

In an open-ended question, we asked if individuals had other comments or thoughts with regards to assessment or feedback for residents on-call. These answers are included in Appendix 4.

2.4 Discussion

The purpose of this survey was to examine the current state of feedback and assessment on-call in Canadian PRS program. We sought to gather responses from both residents and program directors. Eighty-nine percent of all surveyed residents report having five or more call shifts per month (31% having 8-10 shifts, 58% having 5-7 shifts). The larger number of call shifts likely represents PGY2 or 3 residents who typically have higher call requirements compared to more senior years. During these call shifts, most residents see at least 2-4 consults and 86% of residents describe being indirectly supervised on call, with some variation in terms of how often they end up reviewing with staff overnight. These findings are in line with previous studies examining home-call in both Canadian and American settings [53,95]. Time on-call is a substantial portion of training workload as residents, and this is time that almost never involves direct supervision from staff outside of when patient cases go to the operating room. For this thesis, any tool that is developed to assess and provide feedback to residents on-call cannot therefore rely on direct observation, which is recommended as the optimal way to provide feedback[18,58(p20)]. Activities on-call are not amenable to in-the-moment WBAs based on direct observation[22].

In terms of satisfaction with the amount of feedback typically given on-call, 54% of residents were satisfied or very satisfied, whereas 40% were neutral and 6% unsatisfied. Only 39% of residents were satisfied or very satisfied with the quality of feedback. Most residents agreed they would benefit from having a more formal way of receiving feedback. We were surprised by the fact most residents were satisfied with the amount of feedback received. Other studies have shown that, in general in medical education, feedback occurs infrequently and is usually of low quality[96–98]. The fact that 40% and 53% of residents were neutral about the amount and quality of the feedback given, respectively, may relate to known variability in resident engagement with feedback and feedback-seeking behaviours[98,99]. It is clear there is room for improvement in this area, at least for those residents interested in receiving more feedback.

These findings are based on residents working home-call shifts, which is not universal practice. No other studies have examined the provision of feedback in this setting.

Limited reports describing feedback given to residents based on other call systems, like night-float or in-house call exist[55]. We do have a reasonable understanding of how time on-call is spent for residents in surgical or non-surgical programs[52,100–102]. The objectives of many of these studies are to identify resident activities without educational value and find ways to reduce the burden of these activities in the context of duty hour restrictions, with the idea that given limited hours, the proportion of educational to non-educational activities should be maximized. We propose that another approach for maximizing the educational benefit of on-call time is to improve the feedback given to residents. This could be done in addition to working on reducing the burden of non-educational activities.

There were some similarities and differences in the aspects of feedback residents said they would want feedback on and what program directors thought would be important to give feedback on. Both groups agreed that feedback on overall clinical judgement was important. All PDs thought feedback on communication with the medical team was important, whereas only 17% of residents did. Eighty-three percent of residents want feedback on clinical outcomes and 66% on technical-related decisions, compared to 60% and 40% for the PDs, respectively. These differences could reflect the fact the questions posed were not the same for both groups. PDs have a goal of ensuring their program trains residents to become well-rounded plastic surgeons whereas surgical resident may prioritize the technical skills performance. Another explanation could be that PDs, as supervising staff, are on the receiving end of resident handover and communication and so may prioritize this skill.

2.5 Conclusion

This chapter sought to examine the current state of feedback and assessment based on what residents do on-call in Canadian PRS programs through a web-based survey sent to all PGY 2-5 residents and program directors. Based on our results, there is a need for improvement in feedback and interest in a more formal way of providing it. The areas to provide feedback on should be considered from both the training program and staff perspective as well as the resident perspective.

Limitations to this study exist. We did not collect home program as a demographic detail due to the fact this would be enough in some situations to identify individuals (e.g., there is only 1 program director per program and some resident years at a school only have 1 resident). This leads to a potential source of bias given the survey may not represent all programs in Canada. However, the individual response rate was high, with only one resident not completing the entire survey. The overall survey response rate was 30% overall, which is within the typical range of physician responses to web-based surveys[103]. The generalizability of these results to other surgical specialties is also unknown as this study only included PRS programs which typically have a home-call system in place.

We consider this survey to be an important first step in outlining the gap that exists regarding assessment and feedback on-call. First identified through informal discussion in our program, these results show the gap also exists in other PRS programs.

Chapter 3

3 Tool Development and Pilot

3.1 Introduction

Trainee progression in CBME is based on demonstration of competence in real clinical settings. WBAs are commonly used in CBME to test, measure and evaluate competence, and serve as a proxy for what will be done in future independent practice[21]. Other important characteristics of WBAs in CBME are that they should occur frequently, be low-stakes and formative[17,20]. Multiple types of assessments, completed by multiple assessors, should be used by training programs. Ideally, when a program is looking to implement a new WBA, they should use or adapt existing tools that come with validity evidence and are applicable to the construct of interest for the clinical situation[58]. The construct of interest for this thesis is surgical resident competence on-call, and there are currently no existing tools that examine this construct and contain validity evidence. In this chapter, we will use evidence-based assessment design principles to create a formative tool looking at competence on-call. We will pilot the tool within the Division of Plastic and Reconstructive Surgery at our institution and examine tool uptake, use, and perform a quantitative analysis of the resulting scores.

3.2 Methods

We used a checklist for developing a good assessment instrument adapted from Hamstra (2012) to guide our methodology (Table 5)[8].

Table 5. Checklist for assessment development

Step	Description
1	Determine the purpose of your assessment
2	Identify main construct of interest and stakeholders
3	Review with content experts
4	Item writing and tool development
5	Train the raters
6	Pilot test the instrument for validity

Research Ethics Board Approval - Western REB granted approval for this study on March 15, 2022 (Appendix 5).

3.2.1 Determine the purpose of the assessment

Based on early, informal discussion with residents in the Division of PRS at our institution, a desire for a formative assessment of on-call performance was identified. Residents were interested in more feedback based on patient encounters and procedures done on-call. A preliminary tool was developed in 2020 based on the input of three PRS surgeons and one resident (Appendix 6). This tool assessed procedures performed on-call based on the quality of surgical outcomes, surgical adjuncts used, post-procedure plan and patient satisfaction. This tool was not formally piloted. For this thesis, we expanded the purpose and scope of this initial tool to look at more general competence on-call for any surgical specialty, not just limit the tool to assessing technical competence in plastic surgery.

Other a priori specifications were determined for the proposed tool:

- Completion must be possible without relying entirely on direct observation
- Should include both narrative comments and numeric ratings

- Maximum of one-page to optimize feasibility, without a maximum limit on the particular number of items within tool
- Staff surgeons would be the ones completing the tool, but could get input from patients depending on the nature of the suggested items
- Residents would "trigger" the tool to be completed by staff after an encounter on-call
- Each tool completed would be associated with one patient encounter, but responses to some items could represent broader performance over the course of the call shift

3.2.2 Identify main construct of interest and stakeholders

The construct of interest was *surgical resident competence on-call*. A literature review was conducted examining this construct with the help of an experienced medical librarian. The search protocol can be found in Appendix 7.

Several stakeholders were considered in the assessment process. This includes both residents and staff physicians from any surgical specialty; however, the initial scope of the project was limited to residents and staff in PRS, General Surgery and Orthopedic Surgery. Patients were considered another important stakeholder. Given the formative intent of this tool, hospital administration and regulatory bodies were not considered important stakeholders at this stage in tool development.

3.2.3 Review with content experts

To generate content for the construct of interest, we convened a consensus group consisting of experienced surgeon educators and a senior surgical resident. Members were selected using purposive sampling. The purpose of the consensus group was item generation relating to our construct of interest – i.e., we asked group members to consider what were essential components of surgical resident competence on-call. We used the Nominal Group Technique consensus method, because it can be used for item generation and allows for face-to-face group discussion[69]. We followed the methodological recommendations as described by Humphrey-Murto (see Table 6)[69].

Table 6. Checklist for consensus group methodology

Checklist Recommendations	Description of Recommendations	Project Adherence
Define the purpose or objective of the study	Mention if purpose is item generation or ranking or both	Defined in methods
Outline each step in the process	Includes modifications made, provide rationale for each choice	Outlined in methods
Number of participants indicated		Described in methods and results
Selection and preparation of scientific evidence given to participants	Describe what was provided	See Appendix 8 for background information given to participants based on a literature review
Describe how items were selected for inclusion in the initial questionnaire	What protocol was used – idea generation, pre-determined list, etc.	We did not prepare an initial list or questionnaire
Describe how participants were selected and their qualifications	Rationalize the number chosen as well	Selection was based on purposive sampling. Qualifications are described in results

Describe the facilitator's qualifications		Described in methods
Describe number of rounds or criteria for termination	Number of rounds conducted two or more Number of rounds determined a priori	Decided on a maximum of three rounds. Termination criteria described in methods
Describe how consensus was achieved	Description of consensus, polling, use of forced consensus	A priori consensus level determined. Polling described in methods. Did not force consensus
Report response rates after each round and scores	# Respondents indicated	Outlined in results
Describe how anonymity was maintained		Maintained through anonymous survey voting
Describe the type of feedback provided after each round	Formal feedback of ratings to group	Described in methods and results

Adapted from Humphrey-Murto[69]

The steps taken to conduct the consensus group are outlined below:

I. Define the purpose or objective

The purpose of the consensus group was to generate items relating to the construct of interest - *surgical resident competence on-call*.

II. Participants

Participation criteria were: must be "experts" in the field of interest with surgical education experience, must practice in an academic hospital, must (as a group) represent multiple surgical specialties, and must include at least one resident. We used purposive sampling to select six individuals meeting these criteria. This number of participants was based on recommendations that 5-12 group members is ideal for nominal group technique[66]. Five surgeons participated in the group from the divisions of PRS, Orthopedic Surgery and General Surgery. A description of their specialty and qualifications can be found in Table 7. All were actively involved in surgical education. The average academic hospital clinical experience between surgeons was 17 years (starting from the year fellowship was completed including 2021). The consensus group facilitator (EM) was considered a credible non-expert[71]. EM also served as primary researcher for the overall project but did not engage in group discussion or item generation to limit any potential conflict of interest.

Table 7. Description of consensus group members

Consensus Group Member	Specialty	Qualifications
1 – Staff	Plastic & Reconstructive Surgery	Program director
2 – Staff	Plastic & Reconstructive Surgery	CBME lead for division, assistant program director
3 – Staff	Plastic & Reconstructive Surgery	Former program director
4 – Staff	Orthopedic Surgery	Former program director
5 – Staff	General Surgery	Program director
6 – Resident	Plastic & Reconstructive Surgery	Senior resident

III. Description of scientific evidence given to participants

Based on the literature review search described in section 3.2.2., participants were sent a background document with information on the construct of interest and rationale for the consensus group (Appendix 8). Information on the construct of interest revolved around the literature surrounding resident competence in general, and more specific examples of competence assessment in surgery. The document referenced multiple tools currently used in graduate medical education to assess residents across a variety of skills and attributes. A literature review found one study where the on-call competence of ophthalmology residents was assessed, and this was included in the document[57]. Other than this study, the review did not reveal other literature on assessment of on-call competence in specific. The nominal question "*what are the important aspects to include*

on a tool assessing surgical resident competence on-call" was included in the background document.

IV. Describe how items were selected for inclusion in the initial questionnaire

Pre-determined items were not included. The original NGT was described as starting with an open-ended question (nominal question), without an initial list of items, to avoid biasing participants[68]. Item generation was the purpose of our consensus group and so we did not provide an initial list.

V. Described any *a priori* specifications

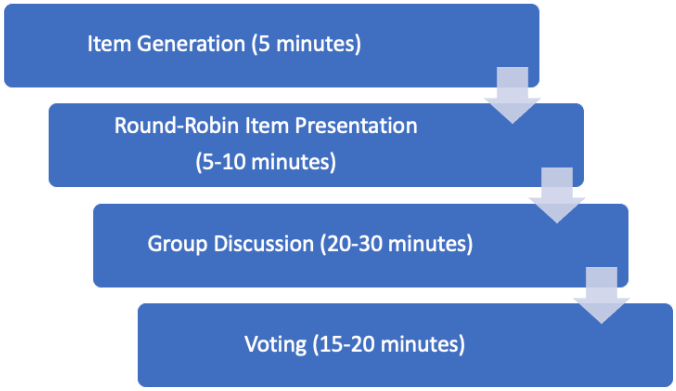
There was a maximum of 3 rounds of voting for consensus. Consensus was defined as at least 80% of participants selecting "agree" or "strongly agree" on a 5-point Likert scale when deciding whether a certain item was important when assessing resident competence on-call. Automatic item exclusion would occur if at least 60% selected "disagree" or "strongly disagree". Items not meeting consensus inclusion or exclusion after the first round of voting were included in a second round of voting with the same criteria applied. If there were greater than 15 consensus items after the end of the second round of voting, a third round would ask each participant to rank their top 10 (10 being most agreeing should be included as consensus, 1 being least agreeing) and the 15 items with the greatest sum of points would be accepted as the final group consensus. Consensus was not forced. Time restrictions were anticipated for the consensus group due to scheduling. As such, it was decided that if the voting could not be done during the meeting itself, an online anonymous voting survey would be sent to each member to complete.

VI. Day of NGT process

On the day of the consensus group, the facilitator provided a brief background on the construct of interest, like information included in the background document and explained the meeting steps to participants (Figure 3). Step one involved five minutes of

silent item generation where participants were asked to consider the nominal question and write down items. Step two involved moving through the group in a round-robin fashion for item generation. The facilitator recorded items as they were suggested on a live document viewable to all members. This step continued until there were no unique items brought up. Step three was open discussion of the items and clarification as needed. Step three concluded with the group having determined their list of items to vote on. Step four involved voting on the items. Each member was asked to rate each item on a Likert scale of one to five, one being "strongly disagree" with the item being important for assessing competency on-call, and five being "strongly agree". Voting was done using an online, anonymous survey (Qualtrics). The results of the survey were passed back to participants by email, as the meeting time had elapsed. The second round of voting was completed in a similar fashion with the results once again passed back to members.

Figure 3. Consensus group meeting steps



3.2.4 Item writing and development

In this stage, we first considered whether there were other tools available that would be suitable to use or adapt. Only when there are no appropriate tools should a new instrument be developed[58]. We reviewed whether any existing tools met our criteria: WBA design, able to be completed without direct observation, applicable to our construct of interest as defined by the consensus group, and it should already have some amount of validity evidence. Tools with good validity evidence such as OSAT, mini-CEX, CAMEO, GEARS, and GOALS are all designed to be completed based on direct observation[28,30,32,104,105]. The OSAT and O-SCORE tools are used specifically for

technical skills assessment in surgery[28,29]. Depending on the results of the consensus group, these tools could be adapted for use with indirect observation. The OCAT has good validity evidence for its use in a surgical clinic and was also considered as an option for adaptation depending on the results of the consensus group[35]. The "On-Call Assessment Tool" developed by Golnik et al. is a one-page tool that assesses "3 critical aspects of on-call performance"[57]. These aspects were identified by a literature review and include patient care, timeliness, and sense of urgency. The study examined the face, content, and discriminative validity of the tool, but did not provide evidence based on modern validity theories and did not examine reliability. This tool is the only tool in the literature identified as being specific to the on-call period, however given it was developed specifically for an ophthalmology program and was not created using evidence-based tool development principles we do not believe it is appropriate for adaptation for use in our study.

EM and AG met formally to discuss the progression of entrustability on call and design a scale with that as a construct. The team members considered the general progression of independence on-call in surgical programs, typically going from direct oversight (often from a senior resident for a junior resident) to indirect oversight with decreasing levels of direct supervision. Other entrustability scales were reviewed from the assessment literature (Table 8) to guide creation of the novel 5-point entrustability scale.

The items from the consensus group were modified for incorporation in the tool and descriptors were developed based on discussion during the consensus group as well as the opinions of the research team.

Table 8. Example construct-aligned scales

Warm et al. [106] 1. Resident not trusted to perform activity even with supervision 2. Resident trusted to perform activity with direct supervision 3. Resident trusted to perform activity with indirect supervision 4. Resident trusted to perform activity independently 5. Resident trusted to perform activity at aspirational level	Kalet et al. [107] 1. Poor – I would not feel safe sharing patient care with this intern 2. Fine – this intern needs very attentive supervision to safely care for patients 3. Satisfactory – this intern can cover my patients with the usual supervision 4. Good – this intern can be trusted to cover my patients 5. Excellent – I trust this intern will provide excellent patient care even when supervision is unavailable
Gofton et al. [29] 1. I had to do 2. I had to talk them through 3. I had to prompt them from time to time 4. I needed to be in the room just in case 5. I did not need to be there	Aylward et al. [109] 1. Cannot perform 2. Can perform under direct supervision 3. Can perform with indirect supervision 4. Can perform independently 5. Can supervise junior trainees
Whalen et al. [108] 1. Direct supervision with supervisor physically present 2. Indirect supervision with direct supervision immediately available 3. Indirect supervision with supervising physician immediately available by telephone/electronically 4. Oversight – supervisor available to provide review of procedures/encounters with feedback after care is provided	Mink et al. [110] 1. Trusted to observe only 2. Trusted to execute with direct supervision and coaching 3. Trusted to execute with indirect supervision with verification afterward for select cases 4. Trusted to execute with indirect supervision with verification afterward for selected complex cases 5. Trusted to execute without supervision

The preliminary assessment tool created was distributed to members of the consensus group for review and feedback on clarity and utility. The tool was revised based on this feedback in an iterative fashion.

3.2.5 Train the raters

PRS faculty were oriented as to the purpose of the tool and how to complete it at a Divisional executive meeting. Residents were similarly oriented as to the purpose of the tool and how to trigger an assessment. It was clearly emphasized to all parties that the tool was formative and that results would not affect academic standing.

3.2.6 Pilot test the instrument for validity

The tool was piloted in the Division of PRS over 6 months (January to June 2022). Residents had three options to trigger an assessment: 1. emailing a copy of the form to the staff, 2. through New Innovations (New Innovations, Uniontown, Ohio), an online resident data management system already in use by Western University, or 3. by requesting completion of a paper copy available in the main clinic spaces of University, Victoria, and St. Joseph's Hospitals. Staff completed the assessment either with the patient as an inpatient or on an outpatient basis. After the form was completed, it was sent back to the resident for review. Anonymized data from completed tools were collated in a password-protected Excel® file kept in an encrypted folder on the hospital network.

During the pilot period, an open channel of communication existed between EM and the participants. Informal feedback was taken into consideration and the process was adjusted as needed. For example, some staff had logistic issues with the New Innovations electronic system, and these were addressed. No changes to the tool were made during the pilot process.

Tool validity was assessed using quantitative and qualitative methods. Descriptive statistics of items were completed including item mean scores, standard deviation, kurtosis, and skew.

We applied generalizability theory to conduct a generalizability study (G study)[111]. The facets, or sources of variation within the model, were resident (r), PGY level (l), occasion (o), and item (i). The G study provided coefficients for internal consistency. G-string software was used to conduct the study[111]. Two D-studies were run, one with item as a random facet and occasion as fixed to look at generalization across items. Another with item as fixed and occasion as random to look at generalization across occasions.

We examined the effect of PGY year on scores using an ANOVA with post-hoc Tukey t-tests. Statistical significance was set at $p<0.05$.

The qualitative validity analysis will be discussed in Chapter 4.

3.3 Results

The results of the tool development process are reported using the same steps outlined in the methods.

3.3.1 Determine the purpose of the assessment

The purpose of the tool was to improve the feedback given to residents based on their patient encounters on-call with a formative intent. Specifically, what on-call competence entails was defined by the consensus group and is outlined in section 3.3.3. The tool was designed to be applicable to surgical specialties outside of PRS.

3.3.2 Identify the main construct of interest and stakeholders

The construct of interest for this tool was *surgical resident competence on-call*. Stakeholders involved in the development and implementation process included surgical residents, staff physicians and patients.

3.3.3 Review with content experts

There were 14 initial items generated during the round-robin for the construct of interest (Table 9).

Table 9. Initial consensus group items generated

Initial Consensus Group Items
1. Communication with patient (explanation of problem, consent process, etc.)
2. Development of rapport with patient (overall bedside manner, professionalism)
3. Time management on call
4. Recognition of urgency of presentation (ability to triage)
5. Handover process
6. Appropriate use of backup / knowing own limitations / knowing when to seek out advice
7. Appropriate follow-up plan (involvement of other services (CCAC), appropriate medications, etc.)
8. Overall management plan
9. Surgical adjuncts used (splints, dressing, etc.)
10. Documentation of encounter (procedure note appropriate, particularly for telephone / virtual encounters)
11. Technical decision-making (suture choice, incision placement, procedural plan)
12. Use of investigations or resources (bloodwork, cultures, imaging, etc.)
13. Clinical-judgement and/or decisions for treatment (e.g., discharge / admission decisions, admission under right service, operative decisions)
14. Procedural outcomes / clinical outcomes

During the discussion portion of the consensus group meeting, multiple salient points were brought up. One was regarding patient input for the assessment tool. Multiple group members agreed that some input from patients is critical for assessing roles like communication and professionalism. However, some staff had previously had trouble with eliciting or utilizing patient feedback. Two reasons posited for this were, firstly, patients may feel that providing a poor rating would negatively impact the care they receive, and secondly, patients may have difficulty differentiating between "poor" communication in a resident and "excellent" communication. Another point made in the discussion was regarding the assessment of technical ability on-call. The staff physicians in the consensus group generally agreed there are multiple other ways to assess technical skills using validated tools, in environments where direct supervision is possible, like the operating room. However, a point was made that residents seem particularly interested in receiving feedback on technical aspects, whether that be on a procedure or choice of surgical adjunct (e.g., splint choice). A final consideration was what an on-call assessment tool could uniquely assess, compared to other established tools. Members generally agreed that the primary unique characteristic was it could provide feedback on multiple aspects of care through the lens of more independent decision making and performance compared to daytime hours.

After the first round of voting, seven items reached consensus and seven items did not (Table 10). No items were automatically excluded. The results of the first round were sent to group members by e-mail. The seven items that did not reach consensus were voted on in a second round. Of these items, three met consensus and four did not (Table 11). These results were sent to group members as well. A total of ten items met consensus through two rounds of voting (Table 12).

Table 10. Round 1 voting results

		Strongly disagree	Disagree	Neutral	Agree	Strongly Agree	Consensus?
1	Communication with patient (explanation of problem, consent process, etc.)	0	0	2	1	3	No. Round 2
2	Development of rapport with patient (overall bedside manner, professionalism)	0	0	1	2	3	Yes
3	Time management on call	0	2	1	0	3	No. Round 2
4	Recognition of urgency of presentation (ability to triage)	1	0	1	1	4	No. Round 2
5	Handover process	0	0	0	0	6	Yes
6	Appropriate use of backup / knowing own limitations / knowing when to seek out advice	0	0	0	0	6	Yes
7	Appropriate follow-up plan (involvement of other services (CCAC), appropriate medications, etc.)	0	0	0	2	4	Yes
8	Overall management plan	0	0	0	3	3	Yes
9	Surgical adjuncts used (splints, dressing, etc.)	1	0	2	3	0	No. Round 2

10	Documentation of encounter (procedure note appropriate, particularly for telephone / virtual encounters)	0	0	0	1	5	Yes
11	Technical decision-making (suture choice, incision placement, procedural plan)	1	0	4	0	1	No. Round 2
12	Use of investigations or resources (bloodwork, cultures, imaging, etc.)	0	0	2	3	1	No. Round 2
13	Clinical-judgement and/or decisions for treatment (e.g., discharge / admission decisions, admission under right service, operative decisions)	0	0	0	1	5	Yes
14	Procedural outcomes / clinical outcomes	2	0	2	1	1	No. Round 2

Table 11. Round 2 voting results

		Strongly disagree	Disagree	Neutral	Agree	Strongly Agree	Consensus ?
1	Communication with patient (explanation of problem, consent process, etc.)	0	0	0	3	3	Yes
2	Time management on call	0	1	0	3	2	Yes
3	Recognition of urgency of presentation (ability to triage)	0	0	0	1	5	Yes
4	Surgical adjuncts used (splints, dressing, etc.)	1	0	2	3	0	No
5	Technical decision-making (suture choice, incision placement, procedural plan)	1	0	2	2	1	No
6	Use of investigations or resources (bloodwork, cultures, imaging, etc.)	0	0	2	2	2	No
7	Procedural outcomes / clinical outcomes	1	0	2	3	0	No

Table 12. Final consensus group items

Final Consensus Group Items
1. Communication with patient (explanation of problem, consent process, etc.)
2. Time management on call
3. Recognition of urgency of presentation (ability to triage)
4. Development of rapport with patient (overall bedside manner, professionalism)
5. Handover process
6. Appropriate use of backup / knowing own limitations / knowing when to seek out advice
7. Appropriate follow-up plan (involvement of other services (CCAC), appropriate medications, etc.)
8. Overall management plan
9. Documentation of encounter (procedure note appropriate, particularly for telephone / virtual encounters)
10. Clinical-judgement and/or decisions for treatment (e.g., discharge / admission decisions, admission under right service, operative decisions)

3.3.4 Item writing and tool development

Based on the consensus group results, we decided to create a new tool, rather than adapt an existing tool. The OSAT, O-SCORE and OCAT were considered for adaptation, however the OSAT and O-SCORE focus primarily on technical skills assessment, which was not a focus in our final consensus group items[28,29,35]. We decided not to adapt the OCAT as it did not have adequate item overlap with the items from our consensus group given it was designed to assess competence in clinic[35].

Item descriptors and tool design

Initial item descriptors or prompts were written during a meeting between the primary researchers (EM and AG) based on discussion during the consensus group. Multiple changes and edits to the items, their descriptors and the tool design were made during further meetings between all study team members, email correspondence and trialing of the tool by AG. An initial tool version can be seen in Figure 4.

Figure 4. Initial tool version

OnCAT – On-Call Assessment Tool

Trainee ID#:	PGY: 1 2 3 4 5	Staff:
Date of Call Shift:	Today's Date:	
Case Complexity: Low Medium High		

The purpose of this assessment is to evaluate resident performance on-call and to provide feedback. With that in mind, please see the scale below to rate each item, irrespective of the resident's year of training. Base your rating on how the resident performed for the specific on-call encounter. Please provide narrative feedback in the spaces below as well.

Items 8 and 9 involve patient feedback, please indicate if you did or did not elicit patient feedback.

Scale

1 **Not able to perform** – i.e., should be left to a senior member to do
2 **Able to perform with direct oversight** – i.e., supervision present in-person
3 **Able to perform with some indirect oversight** – i.e., demonstrates some independence, but requires input and direction
4 **Able to perform with minimal indirect oversight** – i.e., largely independent, but may require assistance in complex, nuanced situations
5 **Competent** – i.e., able to perform independently without direction or guidance

1	**Ability to triage** e.g., recognition of urgency of presentation	1	2	3	4	5
2	**Recognition of need for support or backup** e.g., requests assistance when required	1	2	3	4	5
3	**Follow-up or disposition plan** e.g., involvement of other services, appropriate prescriptions	1	2	3	4	5
4	**Documentation** e.g., consult/admission/procedure/telephone/virtual notes	1	2	3	4	5
5	**Handover** e.g., effective communication with supervising staff/colleagues	1	2	3	4	5
6	**Clinical-judgement and treatment decisions** e.g., decision to discharge/admit/take to operating room	1	2	3	4	5
7	**Overall management plan** e.g., general plan being complete and safe for patient	1	2	3	4	5

Patient feedback elicited Yes No

		Poor	Fair	Good	Very good	Excellent
8	**Development of rapport with patient** e.g., overall bedside manner, professionalism, building trust	1	2	3	4	5
9	**Patient communication** e.g., using respectful and clear language, appropriate consent process and explanation of problem to patient where appropriate	1	2	3	4	5

Global Assessment - Resident is competent and able to perform independently on-call: Yes No

Please provide feedback on what was done well or what could be improved upon:

_____

_____

Staff Signature: _____

Multiple decisions occurred in the design process:

- One of the first decisions made was to use a separate scale for "development of rapport with patient" and "communication with patient". The novel scale we developed for the other items which will be described in further detail, did not

apply as well as these two items were partially dependent on patient feedback. Therefore, a scale adapted from the Communications Assessment Tool was used instead[112].

- Given that patient feedback would not always be obtainable (e.g., ICU patient who is intubated, or staff unable to elicit feedback from patient prior to completing assessment) items "development of rapport with patient" and "communication with patient" were placed in a separate section from the rest of the items and a checkbox was added to select if patient feedback was or was not received. Separating this portion of the tool also emphasized that a different scale was being used for scoring. A comment from one of the team members based on the initial tool draft was "we need to clearly differentiate between the patient feedback scores since they have a different ranking scale. When one fills it out currently the eye is drawn to the 1-5 poor to excellent patient feedback scale when you are looking at [items like] ability to triage or documentation. Therefore, I would suggest either putting the feedback narrative lines in between or using double/triple lines to differentiate the two [sections]".
- The consensus group item "time management on-call" was considered to be similar to the item "ability to triage" and so the two were integrated into one item with the descriptor written as "appropriately recognized urgency and prioritized competing demands on time".
- There was initially a box at the top of the form asking for a rating of case complexity – low, medium or high. A comment by one of the team members was that it seemed like it could be very subjective or arbitrary to judge. The other team members agreed with the recommendation to remove it.
- Instead of the case complexity box as seen in the initial tool version (Figure 4), we added a box for "patient presentation". A space to fill in the patient presentation could serve as a reminder for staff as to which case from on-call they were providing feedback on.
- A decision was made to not include a global assessment item at the end of the tool because it was confusing to readers as to its purpose.

Scale design

The initial 5-point entrustability scale developed by AG and EM is seen below in Figure 5.

Figure 5. Initial scale version

Scale

1. **Not able to perform** – i.e., should be left to a senior member to do
2. **Able to perform with direct oversight** – i.e., supervision present in-person
3. **Able to perform with some indirect oversight** – i.e., demonstrates some independence, but requires input and direction
4. **Able to perform with minimal indirect oversight** – i.e., largely independent, but may require assistance in complex, nuanced situations
5. **Competent** – i.e., able to perform independently without direction or guidance

Other team members provided recommendations to improve clarity but agreed that the 5-points accurately reflected the natural progression of entrustment on call. One specific edit was to change "able to perform independently" to "could have been completed independently". This was because of the potential for a surgeon completing the assessment to avoid selecting that option because it insinuates a lack of supervision. The final scale is seen below in Figure 6.

Figure 6. Final scale version

Scale

1. **Assessment/task incomplete** – i.e., required complete takeover by a more senior physician
2. **Able to perform with direct oversight** – i.e., trainee required in-person supervision to safely complete task
3. **Able to perform with indirect oversight** – i.e., demonstrated some independence, but required verbal or electronic input and direction
4. **Able to perform with minimal indirect oversight** – i.e., task could have been completed independently, but could benefit from direction in complex or nuanced situations for improved outcomes
5. **Competent** – i.e., task could have been completed independently and efficiently without direction or guidance

3.3.5 Train the raters

Staff orientation was done at the beginning of a virtual meeting. The project and tool were explained by AG. The entrustability scale was not specifically reviewed during this meeting. An additional follow-up email was sent explaining the project and tool to all staff members.

Residents were oriented to the purpose of the project and the process for triggering the tool during their weekly academic half-day. They were reminded there was no academic obligation to trigger completion of the tool and that the purpose of the tool was to provide formative feedback. An email was sent to all residents explaining the triggering process in more detail and a copy of this can be seen in Appendix 9.

3.3.6 Pilot test the instrument for validity

The revised tool used in the pilot is seen in Figure 7.

Figure 7. Final tool version to pilot

LOCAT – London On-Call Assessment Tool

Trainee Name:	PGY: 1 2 3 4 5	Staff:
Date of Call Shift:	Today's Date:	
Patient Presentation:		

The purpose of this assessment is to evaluate resident performance on-call and to provide feedback. For the specific on-call encounter, please complete the sections below by using the medical record, direct review with the resident and/or discussion with the patient. Please provide narrative feedback in the spaces below. Items 8 and 9 involve patient feedback, please indicate if applicable.

Scale

1 **Assessment/task incomplete** – i.e., required complete takeover by a more senior physician
2 **Able to perform with direct oversight** – i.e., trainee required in-person supervision to safely complete task
3 **Able to perform with indirect oversight** – i.e., demonstrated some independence, but required verbal or electronic input and direction
4 **Able to perform with minimal indirect oversight** – i.e., task could have been completed independently, but could benefit from direction in complex or nuanced situations for improved outcomes
5 **Competent** – i.e., task could have been completed independently and efficiently without direction or guidance

1	**Ability to triage** e.g., appropriately recognized urgency and prioritized competing demands on time	1	2	3	4	5
2	**Recognition of need for support or backup** e.g., requests assistance when required	1	2	3	4	5
3	**Follow-up or disposition plan** e.g., involvement of other services, appropriate prescriptions, appropriate follow-up	1	2	3	4	5
4	**Documentation** e.g., consult/admission/procedure/telephone/virtual notes	1	2	3	4	5
5	**Handover** e.g., effective communication with supervising staff, transfer of care with colleagues when required	1	2	3	4	5
6	**Clinical-judgement and treatment decisions** e.g., appropriate treatment, decision to discharge/admit/take to operating room	1	2	3	4	5
7	**Overall management plan** e.g., general plan being complete and safe for patient	1	2	3	4	5

Please provide feedback on what was done well or what could be improved upon:

_____

_____

Patient feedback elicited Yes No

		Poor	Fair	Good	Very good	Excellent
8	**Development of rapport with patient** e.g., overall bedside manner, professionalism, building trust	1	2	3	4	5
9	**Patient communication** e.g., using respectful and clear language, appropriate consent process and explanation of problem to patient where appropriate	1	2	3	4	5

Staff Signature: _____

Sixty-three assessments were completed for 10 residents by 7 staff members. The average number of tools completed per resident was 6.3 (range 3-11). Resident participant breakdown based on level of training was three PGY2, two PGY3, two PGY4 and three

PGY5. Twenty assessments (32%) had the optional items 8 and 9 completed based on patient feedback. Fifty-eight encounters (92%) had narrative comments written. There were no forms excluded due to incomplete information or improper scoring. There was a wide variety of patient presentations for which the tool was completed, with representative examples outlined in Table 13.

Table 13. Sample patient presentations

Facial fracture
Tendon laceration
Scalp defect
Flexor tenosynovitis
Hand fracture
Burn
Lip laceration
Leg infection
Nerve palsy
Infected pressure injury

Item Analysis

Table 14 contains the item level descriptive statistics. The average score across all items and all residents was 4.03 ± 0.83. The minimum score given for five of the items (items 2, 3, 5, 6, 7) was 2 and 3 for the other items. No scores of 1 (i.e., assessment/task incomplete – i.e., required complete takeover by a more senior physician) were given for any item. The highest average scores were for patient feedback items 8 and 9 (4.20 and 4.30 respectively).

Five items had significant skew (asymmetry of the distribution of scores around the mean), as signified by a z-score > +/- 1.96. Three items had significant kurtosis (a measure of the shape of a distribution). All items had a negative skew and kurtosis.

Table 14. Item level descriptive statistics for tool pilot

Item #	Average	St Dev	Min	Max	Skew	SEsk	Z-score	Kurtosis	SEk	Z-score
1	4.14	0.74	3	5	-0.23	0.17	-1.41	-1.10	0.39	**-2.82**
2	4.11	0.81	2	5	-0.59	0.17	**-3.55**	-0.19	0.39	-0.47
3	3.90	0.87	2	5	-0.41	0.17	**-2.46**	-0.50	0.39	-1.29
4	4.06	0.74	3	5	-0.10	0.17	-0.61	-1.12	0.39	**-2.87**
5	4.03	0.86	2	5	-0.22	0.17	-1.32	-1.25	0.39	**-3.18**
6	3.92	0.92	2	5	-0.48	0.17	**-2.88**	-0.59	0.39	-1.50
7	3.89	0.94	2	5	-0.38	0.17	**-2.31**	-0.77	0.39	-1.96
8	4.20	0.62	3	5	-0.12	0.17	-0.72	-0.21	0.39	-0.53
9	4.30	0.66	3	5	-0.40	0.17	**-2.39**	-0.55	0.39	-1.40

St Dev – standard deviation, Min – minimum, Max – maximum

SEsk – standard error skew, SEk – standard error kurtosis

Z-score = Skew / SEskew

Bold: significant Z-score > +/- 1.96

Generalizability Analysis

A G-study was conducted, with facets of resident (r), PGY level (l), occasion (o), and item (i). Nested or crossed facets included resident nested in level (r:l), occasion nested in resident and level (o:r:l), level crossed with item (li), resident crossed with item nested in level (ri:l), and occasion crossed with item nested in resident and level (oi:r:l). These facets are explained in Table 15. The scores from items 8 and 9 were not included as

these were not completed for each tool occasion. Table 16 shows the results of the G-study in terms of variance components for each facet.

Table 15. G-study facets and explanation

Facet	Explanation
PGY level (l)	The variance attributable to resident year of training
Item (i)	The variance attributable to the items of the tool
Residents within level (r:l)	The variance attributable to the resident differences in a certain PGY level
Occasion within residents within level (o:r:l)	The variance attributable to the different occasions of scoring for a certain resident in a certain PGY level (e.g., between occasion 1 and occasion 2 of tool completion for Resident A in PGY3)
Level crossed with item (li)	The variance attributable to PGY level and individual items (crossed because all items are assessed within all PGY levels)
Residents crossed with item within level (ri:l)	The variance attributable to individual residents within a PGY level looking at individual items (e.g., item ratings for resident A in PGY3 and resident B in PGY3)
Occasion crossed with items within residents within level (oi:r:l)	The variance attributable to individual residents within a PGY level looking at

	individual items and all occasions, plus random error

Table 16. G-study variance results

Facet	SS	MS	VC	% of Variance
l	132.34	44.11	0.36955	44.64
r:l	30.72	5.12	0.08405	10.15
o:r:l	80.91	1.53	0.19432	23.47
i	4.09	0.68	0.00592	0.72
li	5.11	0.28	0.00801	0.97
ri:l	5.91	0.16	-0.00034	-0.04
oi:r:l	52.88	0.16	0.16630	20.09

Facets - l : level of training, r : resident, o : occasion, i : item

SS - sum of squares, MS - mean squares, VC - variance component

44.64% of the variance came from resident level of training. Variability among occasions for an individual resident in a given year of training accounted for the second most variance of 23.47%. Variability attributed to the items within the tool was minimal (facets – i, li, ri:l). The variance components in Table 16 produced a generalizability coefficient of 0.67, which represents overall tool reliability.

Two decision-studies were done. The first D-study used occasion as a fixed factor and item as a random factor, to look at generalization across items. This gave a generalizability coefficient of 0.92. This value is equivalent to internal consistency in CTT. The second D-study used occasion as a random factor and item as a fixed factor, to

look at generalization across occasion. This gave a generalizability coefficient of 0.28. This represents the reliability of a single occasion of scoring.

Relation to PGY-Year

The tool completion and mean score breakdown (excluding items 8 and 9) based on PGY-year is seen in Table 17.

Table 17. Score by PGY-level

Resident Year (PGY)	Mean Score	St Dev	# Forms Completed
2	3.53	0.67	26
3	3.74	0.52	14
4	4.37	0.49	10
5	4.93	0.25	5

Year of training had a significant effect on mean scores – $F(3,59) = 23.34$, $p<0.001$. This analysis excluded items 8 and 9 which were not completed in each case.

Post-hoc Tukey HSD tests revealed a significant difference comparing PGY2 to PGY4 scores ($p<0.001$) and PGY2 to PGY5 ($p<0.001$) scores. PGY3 scores were significantly lower than PGY4 ($p<0.05$) and PGY5 ($p<0.001$) scores. There was no significant difference between PGY2 and PGY3 scores or PGY4 and PGY5 scores.

When just examining items 8 and 9, there was a significant effect of PGY-year on mean scores – $F(3,16) = 6.63$, $p<0.01$. Post-hoc tests showed PGY2 patient scores (mean 3.75±0.61) were not significantly different from PGY3 (4.00±0.00) but were significantly different from PGY4 (5.00±0.00, $p<0.05$) and PGY5 (4.64±0.48, $p<0.05$). There were no significant differences in comparisons between PGY3, 4, or 5 scores. The number of forms completed with patient scales complete were overall low (PGY2 = 6, PGY3 = 5, PGY4 = 2, PGY5 = 7).

3.4 Discussion

We used an adapted checklist to develop our novel assessment tool[8]. Adhering to a systematic process of tool development is critical to creating an assessment that aligns with its intended purpose and construct of interest. Testing the tool in a real clinical setting, collecting validity evidence, and measuring reliability is part of the development process, as is making necessary revisions and modifications. In this study we laid out each step, reported our methodology and results to ensure transparency.

The purpose of our tool was to provide feedback given to residents based on their patient encounters on-call. This had a formative intent and originated from requests in our division for more feedback based on what was done on-call. Residents spend a significant portion of their residency on-call, but despite this, a review of the literature failed to identify consensus on what on-call competence entails or how, if at all, it may differ from competence displayed during the day. Ideally, we would have adapted a tool with existing validity evidence to assess our construct of interest; however, no tools were deemed relevant enough to do so. Therefore, we used the consensus group methodology nominal group technique to outline what surgical resident competence on-call entails. We believe inclusion of staff surgeons from specialties outside of the division of PRS helped us build a tool which can be applied more broadly to other surgical specialties, a secondary goal of our tool development process. Whether this goal was truly achieved requires testing of this tool within other specialties.

The results of our consensus group show that competence on-call involves multiple CanMEDs roles, with a tendency toward non-technical skills. Consensus group members did not think that technical-related decisions and outcomes were important to assess on-call. This decision may be related to technical-related skills being more easily and appropriately assessed directly in other settings. Technical skills are certainly important in surgical residents who are doing procedures overnight or taking cases to the operating room. A tool to assess technical skills, however, may not be as critical as staff would be providing direct supervision for operative cases and therefore could provide more direct feedback as needed. Non-technical skills have traditionally received less focus in surgical residency training, despite having a significant impact on patient care[113–115]. The

importance of strong non-technical skills in residents may be heightened on-call, when staff are less present, and residents must effectively communicate, manage their time, and make clinical decisions independently. It would be helpful to further investigate the reasoning behind the focus on non-technical skills in future studies.

There are some similarities in these results to internal medicine literature in terms of the core skills emphasized on-call[56,116]. Competence on-call clearly requires functioning across multiple roles and understanding and assessing these aspects may allow residency programs to better prepare their residents for being on-call not only during residency but in practice as well. Having a means to assess these aspects could help guide decisions on when to transition a resident from a junior to senior level of responsibility, which is a large step within training[56,116,117].

Through a generalizability study, we found the largest sources of variance in our tool were from differences among PGY levels and different occasions of an assessment being completed for a certain resident in a specific PGY level. The high variance contribution of PGY level shows us this is an important factor in score differences seen, which is to be expected as, in theory, resident performance should improve over time. We did not include resident as a stand-alone facet in the G-study; however, it would be expected to attribute 44-55% of the variance seen (variance of "l" facet plus variance of "r:l" facet). Variance between individual residents is expected, as performance not only varies between PGY levels but also in residents within a certain PGY level.

The second D-study showed a low generalizability coefficient of 0.28, meaning that the reliability of scores for one assessment occasion for a specific resident compared to scores from another occasion was low. This is somewhat akin to test-re-test reliability. Truly testing test-re-test reliability in our study would involve the same resident completing an assessment on a similar patient presentation of similar complexity with the same staff rater, so the analogy is not completely valid. There are many factors that change from one assessment occasion to the next which explain the low reliability.

The overall reliability of our tool was 0.67. For the pilot results of a formative assessment tool with limited number of tool completions, this is within what would be expected. We

did find a high coefficient of 0.92 in the D-study looking at generalization between items. This suggests our items all closely align with each other, a measure of internal consistency. Overall, these results can be used to tell us that to improve overall reliability, the number of occasions of tool completion should be increased. Increasing or decreasing the number of tool items would not have a large effect on improving reliability.

The tool was able to differentiate between more junior (PGY2) and senior residents (PGY4 and 5). The inability of the tool to differentiate between PGY2 and PGY3 residents was of interest. The average score of items one to seven for PGY2 residents was 3.53 +/- 0.67 compared to 3.74 +/- 0.52 for PGY3. Given the small sample size of residents with only three PGY2 residents and two PGY3, it may be that the sample size was too small to demonstrate a difference that does exist. Seeing if there is a correlation between on-call tool scores and other variables likes EPA scores or ITER scores would be helpful to sort this out. This finding could also mean that the largest improvement in competence on-call occurs during late-PGY3 to late-PGY4 period. Further investigation into this will be important. The portion of the academic year that this study was conducted meant that PGY2 residents were already taking independent “senior” call (i.e., without a more senior resident on-call with them). It would be valuable to see whether the scores of residents earlier in their PGY2 year would be different from scores later in the year after the transition to taking senior call had occurred. If this is the case, the tool could be used by program directors to help make the decision of when to transition residents to senior call.

3.5 Conclusions

In summary, we applied consensus group technique to determine the critical features of surgical resident competence on-call, which were primarily comprised of non-technical skills. We developed a formative assessment tool to examine competence on-call and piloted it over a six-month period in the Division of PRS at one institution. Overall tool reliability was 0.67, which could be improved through more occasions of tool completion. The tool was able to differentiate between residents of different PGY-years.

This study has limitations. With regards to the working group, it was led by EM, who does not have experience facilitating this type of activity. We did adhere to well-established guidelines set out by Humphrey-Murto et al. to minimize the impact of an inexperienced facilitator[69]. Another limitation is the tool was piloted in one division at a single institution. Piloting of this tool within other surgical specialties and at other institutions is needed to examine whether it can be generalized. Although the tool was designed based on input from multiple surgical specialties, consensus group methodology should be not used in place of rigorous empirical evidence, but it is a valuable first step in the process of examining the construct of interest[67]. A final limitation is that raters in our study were unblinded to the resident they were scoring. It is possible scoring decisions were made partially based on level of training as well as pre-existing opinions about specific resident competence instead of strictly based on the construct-aligned scale, an example of the halo effect[118,119]. It would be difficult to completely blind raters to the residents they are assessing and, although it is a different tool, the O-SCORE which uses a similar construct aligned scale has been shown to have accurate and reproducible results when used in a blinded vs unblinded fashion[120].

The results of this chapter will be further discussed in Chapter 5 in the context of validity evidence.

Chapter 4

4 Qualitative Investigation of On-Call Feedback and Tool Impact

4.1 Introduction

Assessment plays a critical role in postgraduate medical training. The characteristics of a good assessment vary depending on the type or purpose of the assessment and the stakeholders' needs. Validity is considered by many to be the most important characteristic of assessment data. Other necessary characteristics include acceptability, feasibility, and educational effect. Training programs must decide which tools they will use for the assessment of their residents and, to do so, should understand the nuances of each tool to optimize educational outcomes. Previous studies have used semi-structured interviews to assess some of these characteristics[121–123]. The purpose of this portion of our study was to examine the impact, feasibility, acceptability of the tool we designed, as well as to collect validity evidence, through thematic analysis of semi-structured interviews with residents and staff. We also wanted to examine the current climate of feedback on-call to better understand any changes that occurred because of tool implementation.

4.2 Methods

We utilized purposive sampling of residents and staff from the Division of PRS who triggered (resident) or completed (staff supervisors) at least one assessment tool in the pilot. These individuals were invited by email to participate in semi-structured, single or group interviews. The semi-structured interview guide was developed by EM and AG and can be found in Appendix 10. The interview guide questions were adjusted as needed after each interview. A single interviewer (EM) conducted each interview through videoconferencing. Interviews were recorded, professionally transcribed, and anonymized. Interviews lasted 27 minutes on average. Transcripts were checked against the recordings for accuracy.

Research Ethics Board Approval - Western REB granted approval for this study on March 15, 2022 (Appendix 5).

Thematic Analysis

We utilized the six steps of thematic analysis (TA) described by Braun and Clark to analyze our data[124]. TA is "a method for identifying, analyzing and reporting patterns (themes) within data"[124]. We decided to apply reflexive TA where coding is open and organic, not requiring a codebook or coding framework. TA provides high flexibility, without requiring the highly technical knowledge of other analysis methods, and is useful for understanding the experiences or thoughts of multiple participants[125].

An inductive approach was adopted where the themes emerged from the data rather than being decided beforehand[124,126]. The first two interviews were read independently by EM and AG and systematic coding was done. Coding results were compared and discussed by EM and AG to agree upon preliminary codes and any discrepancies were resolved. The codes were revised iteratively with subsequent interview analysis. EM and AG met regularly to discuss the themes identified in the interviews. Saturation of themes is not a criterion in thematic analysis; however, codes became recurrent and overlapping by interview number seven, and so this was considered the endpoint. This represented 40% of those who were eligible for participation based on our inclusion criteria. NVivo 12 (QSR) was used for coding and analysis. Memos were used to keep track of researcher thoughts and observations during the analysis period.

Reflexivity statement

AG is a practicing academic plastic surgeon and the program director of the PRS program from which residents were recruited. He viewed the data through his clinical teaching and assessment lens as a supervisor for residents on-call, as program director, and as someone with a research background in medical education. He does have influence over the residents in his role as program director; however, he did not participate in the interviews. EM is a resident colleague of the resident participants in this study and a trainee under the staff physicians who participated. EM conducted the interviews and had no power or

authority over the participants. He viewed the data through the lens of a trainee without a significant background in medical education.

4.3 Results

A total of seven participants (three staff supervisors, four resident physicians) were interviewed. One of the resident physicians interviewed was part of the full-time CBME curriculum cohort. Residents were from PGY-2, PGY-4, and PGY-5 levels. The interviews lasted 27 (range 18 to 38) minutes on average

The overall findings will be presented in relation to four overarching themes - baseline feedback on-call, consequences of tool implementation, mediators of tool utility, and suggestions for future directions.

Baseline Feedback On-Call

Limited Amount of Feedback Given

All residents interviewed identified infrequent feedback given when asked to describe their typical experiences when on-call: “Now [that I] think about it, probably not a lot of formal [feedback] unless it was unique.” The explanations for the limited feedback varied: “… maybe [staff will] realize that you did the procedure, but they forget to… follow up and let you know. Especially in a busy clinic, they're not going to be able to do that for all residents.” Other rationales included a recognition that case discussions are generally limited at late hours of the night and that staff supervisors potentially give limited feedback when the case presentation is less related to their area of expertise. This reflection was also shared by one staff participant who explained how forgetting to initiate feedback can happen: “…you know, you don't see the residents, you never give them the feedback or maybe you don't even think about [feedback] because, you know, things are moving so quickly”.

Immediate Case Discussion

Of the feedback that did occur regarding on-call encounters, the majority came in the form of immediate case discussion. Depending on the clinical problem, a resident may or

may not immediately review the case with their supervising staff. Immediate case discussion was usually more common earlier on in training: "…when you're starting to do like senior call… I would say, for… admissions and stuff like that, the staff are generally pretty good at giving immediate feedback… over text." Another resident said: "… when you're talking to the staff [overnight] you are getting immediate feedback on your plan."

This immediate case review conversation most often occurred through text messaging or phone calls. Staff and residents both commented on how case discussion through text messaging is typically brief, composed of staff confirming the plan laid out by the resident with a text response like "Okay, sounds good." If a resident thought a patient needed a procedure done, they described how they send the staff a picture of the injury, with annotations showing how they would approach the repair or incision, for example. In situations such as this, some staff made recommendations on the procedural approach or technique. One staff who was interviewed said they preferred to talk on the phone if it was at a reasonable hour on-call: "…and I think one of the barriers now is technology. With texting, I say this to residents a lot, you get more out if you have an actual [phone] conversation than a text conversation."

Resident Feedback-Seeking

Residents explained that to get feedback outside of that given during immediate case discussion, they would have to seek it out themselves. The type of information they were looking to gain was mostly on patient outcomes and if anything should have been done differently in terms of their decision-making at the time of the consult. Some residents put the onus of feedback seeking on themselves: "… I think [getting feedback] is a resident role to some extent. Like if you want to know how stuff worked out, you should follow up on it. That's what I've done."

Seeking out feedback could take the form of reaching out to staff directly: "Once in a while, especially when I was first starting, I would text them… when I knew [the patient was] coming back to clinic and ask how they were doing. And so, they would kind of text me then, but I would… have to seek it out a little bit. Otherwise, I found there wasn't too much feedback." Another way of seeking feedback entailed following patient notes on

the electronic medical record: “I have a list of people I'm interested in, and I don't always ask, but you see what's going on. You read the notes about them, follow along.” There was variation among residents in terms of the extent to which they sought out feedback.

Consequences of Tool Implementation

All staff and residents thought the tool was generally valuable and useful. Given the lack of feedback at baseline, any improvement in feedback was considered beneficial. The responses revealed many factors playing a role in the quality and quantity of feedback such as, time of night, supervisor expertise, and type of patient presentation. There were no themes relating to negative consequences created because of the tool. A concern brought up by two residents was the potential for increased administrative burden on staff who are asked to complete the tool; however, this concern was not mentioned by any of the staff interviewed as being a problem.

The consequences of tool implementation, as described by interviewees, can be categorized into three sub-themes: 1) quantity of feedback 2) standardization of feedback and 3) type of feedback.

Quantity of Feedback

Both residents and staff said the tool positively influenced the amount of feedback given. One staff commented on how “… anytime we do an evaluation… you end up giving more feedback…”. Even though the feedback may be indirect, residents felt “… any type of response from the staff is super helpful.” The absolute change in the quantity of feedback would be determined by the frequency of tool triggering by each individual resident. This change was especially apparent given the limited amount of baseline feedback that was given.

Standardization of Feedback

A challenge with feedback related to the on-call period is that “… the nature of call and what's being assessed [on-call] isn't always conducive to… feedback in the moment. So having [a tool] … available is… good.” At baseline, staff explained how it is easy to

default to a limited response when reviewing with a resident on call. One staff member said a typical response after a resident explains their plan for care might be to text, "… sounds like a good plan" and that as a staff you "…may never tell them how that patient was in clinic unless you, you know, you see them, and the resident is there." The feedback tool had a positive effect in terms of promoting more reflection on resident performance to inform the feedback given. A staff member said it gave them "… a chance to take reflection and pause.".

This ties into some of the rationales proposed by interviewees for the infrequent baseline feedback given to residents based on what they do on-call. Sending a supervising staff member the feedback tool to complete not only reminds them that this was a patient the resident saw on-call, but also prompts staff to think about what aspects of care or decision-making could be improved upon: "So, I think [receiving the tool] gives you a chance to actually sit down and think, okay, did the resident… go through each of those items [outlined in the tool] and think about that and actually be able to give more constructive feedback. Like you did the assessment really well, but you can elaborate on your plan, or you didn't give the patient instructions." The scored items in the tool provided a structured way for feedback to be given and staff had to think of each item individually, in addition to global performance. Another staff member commented on this aspect as well: "I think that having it standardized allows for feedback that probably I wouldn't give in such a granular way to trainees." Residents agreed the tool formalized and streamlined the feedback process but did not specifically comment on how the feedback within the tool differed from the feedback they would get talking to the staff by phone on-call.

Type of Feedback

Another change to feedback attributed to tool implementation was it elicited feedback on different aspects of the care pathway outside of what typically would be given during immediate case review. Patient feedback was one of these changes. The nature of being on-call, particularly in the program of the participants, is that a patient may be seen in the middle of the night and a resident may not see them again or hear about their follow-up.

This means a resident would not routinely get feedback on communication or rapport building with patients when they come back to clinic.

One resident commented on how the tool is "…good for really having that patient input because I think, at least the ones that I had done by some of the staff, it sounded like they had talked to the patients about me, which I think is good from a… style perspective and learning how to interact with people." Adding in a patient feedback section to the tool was based on the working-group consensus that patient rapport-building and communication were important aspects of on-call performance for a resident. Staff participants did not touch on how or when they approached patients for feedback to include in the tool.

Another consequence of not seeing patients in follow-up is you often do not get feedback based on patient outcomes. One resident commented how the follow-up on things such as how wounds healed, or the appropriateness of antibiotic choices, were often not communicated and completion of the tool was thought to "…[have] a lot of value for that".

Mediators of Tool Utility

There was discussion in the interviews about features that improved or detracted from the feedback captured within the tool or the overall tool triggering and completion process. Part of the discussion had to do with the ideal application of the tool and the other aspect was ease of use. We will discuss these as two separate sub-themes.

Ideal Tool Application

Among residents, there was disagreement about whether the tool was useful for receiving feedback on procedures done on-call. One resident said the "… tool probably has more utility for procedures versus just like regular consults and admissions." They commented how they sent out less tool completion requests for patient encounters that did not have associated procedures. A different resident said that for procedures "… someone needs to be there or available with you…", to provide feedback and coaching in the moment. They described how for suturing and other technical skills "… it's hard to give feedback…

after the fact." This same resident thought that the tool was more useful for commenting on decision making in general, compared to technical skills. Part of the difficulty with technical feedback would be that if a staff member sees a patient in clinic after a resident closed a laceration on-call, they may only be able to comment on aspects of care like appropriate suture choice, or how a repair looks after two weeks, but not aspects like tissue handling, efficiency of repair, suturing technique, etc.

In terms of utility based on level of training, all residents agreed the tool would be most useful to a resident who was starting senior call. In this training program, a resident starts senior call mid-way through their second year, meaning they are no longer doing "junior" or "buddy" call with a senior resident on-call alongside them. As a junior resident on "buddy call", feedback would typically be direct, in-person feedback from the senior resident and so the tool might not be as necessary in that case to facilitate feedback. Additionally, many of the on-call tasks represented by tool items, like making triage decisions, might be done by the more senior resident instead of the junior resident. One resident commented how the tool "…would be very helpful in third year when you're… doing procedures on your own that you've never done before…", more-so compared to fourth or fifth years of training when you are more comfortable with the routine procedures seen on-call. Another resident said for a "[PGY-1 or PGY-2] in the first couple of months, maybe [the tool is] not as relevant. But I think if you are… just starting… call by yourself as a [PGY-3] or… having more responsibilities as a [PGY-4] I think [the tool] will be much more useful and relevant to that."

Staff found it was best to complete the feedback tool in cases where there was some amount of direct interaction with the resident. For example, filling out the tool as staff based on a patient encounter reviewed by phone overnight or the subsequent day was easier to do than if the resident sent the tool off a few days after the on-call night or did not review at all with the staff. Having some amount of direct interaction provided more information for the staff to base the assessment on and prompted staff to think about what feedback to give earlier on. One staff member considered completion of the tool as part of a debrief "the ideal situation", because if "you're actually just talking to someone…"

as opposed to having no interaction with the resident, helpful feedback "… tends to come a bit more.".

Finally, two interviewees brought up the notion that the tool may be most useful for cases that did not go as well or had room for improvement: "…if there's something that's not gone well… then that feedback is probably even more, maybe not valuable, but critical I guess to improvement." There was concern that if the tool was designed to be more summative than formative in nature, residents would only send it out when an interaction or case went well. If an encounter with a patient went perfectly, there would not be as much room for constructive feedback to be given and this could affect the overall tool utility.

Ease of Use

Interviewees commented on both the process of triggering and completing the tool as well as tool design and content. Everyone agreed that the process of triggering and completing the tool was easily done. The fact there were multiple formats of the assessment available was useful as some staff preferred one format to another: "The New Innovations one is certainly easier to follow through a link. And not that the PDF one was overly arduous or anything, but I think … having it in a ready to use format… definitely takes away some of the roadblocks." A different staff member found it was "quicker just being emailed [the PDF] directly than having to log in to a third-party platform…"

The length of the tool was well received as all three staff members liked how it was limited to one-page. The individual item prompts and associated rating scale were considered clear as well: "I think that everybody involved in medical education could follow… [the tool] in a relatively easy way without any teaching or, you know, prompting…and needing really to be briefed on how to do it. It's all it's all… pretty straightforward." Furthermore, regarding the items, one staff member said, "I don't think I ever completed one thinking that there was a category left out or that I wasn't able to adequately explain." When asked about whether they could differentiate between different numbers on the rating scale, staff did think this was clear: "I think I could get a sense of the difference... between a two and a three and a three and a four on the rating

scale, whereas like five is fairly independent. So yeah, I thought that was clear." Another staff said that the rating scale "did seem easy to triage" and that people are generally used to Likert-style scales.

Future Directions

Despite being considered a relatively easy process to trigger and complete a tool, both staff and residents identified potential to improve the experience. We categorized these suggestions under the theme of future directions.

Three residents brought up how it would be helpful for the process to be more automated overall. This may be more relevant for the residents in the CBME/CBD curriculum: "It would be better if [the assessment] was automatic somehow. But I think overall, especially like I'm not in CBD, so it doesn't really matter to me as much, but I think for you guys it will help you get EPAs." Another resident said: "I guess it would be nice if in the future… this was kind of an automatic thing where the resident isn't responsible for triggering…" More automation could also help with reminding staff members of the request to complete the tool. A feature was suggested where you could enter the date the patient you saw on call is being seen for follow-up in and the system would "automatically trigger a reminder or something or just send [the form] out the day before [the visit date]." This would rely on the resident knowing when the patient was coming back and does put the onus on the resident in that regard. Another suggestion was to integrate into a phone app: "Something like [an app] would make [the process] easier." Instead of having to log-in through a webpage or partially complete a PDF to then send to a staff member, one resident said a more streamlined process would be to "have an app on my phone, [where you could] right click here and type in a name and it gets sent."

4.4 Discussion

Based on the results of the semi-structured interviews, there is limited feedback given to residents about on-call performance. When feedback is given, it is largely in the form of immediate case discussion, and the primary driver of the feedback exchange is the resident. Multiple other studies have found low quantity and quality of feedback in other

postgraduate medical education settings[96,97(p),98]. This finding is not surprising, as there are numerous barriers to requesting and receiving feedback overnight including waking staff up from sleep, communication indirectly through text or on the phone instead of in person, time constraints on a busy shift, fear of negative feedback, etc. Patients who come to hospital overnight may have presentations that differ from what the on-call staff typically manages in their day-to-day practice, potentially resulting in lower staff credibility, which has been shown to negatively affect feedback seeking by residents[127]. An example of this could be a sub-specialty reconstructive breast surgeon providing guidance to a resident on a complex craniofacial case on-call. There is clearly room for improving the feedback given based on what residents do overnight on-call.

The results of the national survey in Chapter 2 revealed only a minority of resident were truly unsatisfied with the quantity or quality of feedback given to them on-call, but many residents did express a neutral opinion. In our interviews, although residents clearly outlined the limited level of feedback given to them based on what they did on-call, they did not describe associated dissatisfaction or frustration. However, like what was seen in the national survey results, resident interviewees did express an interest in improving the feedback culture.

Based on our results, implementing the tool appeared to increase feedback provision. Part of this change came from formalizing and standardizing the feedback process. A common reason for a lack of trainee feedback is a poor culture of feedback[98,128]. Delva et al. described how a culture or learning climate that normalizes feedback can make residents seek more feedback[98]. In this study, standardization of the feedback process through use of the tool not only prompted residents to consider requesting feedback, but also forced staff to “sit down and think” about a resident’s performance in more reflective fashion. An emphasis on specific, actionable, and timely feedback has been highlighted in numerous reports in the literature[99,128,129]. Implementation of the tool may play a role in changing the culture of feedback. Further observation is needed to determine whether a tool like this has a long-standing impact on feedback or a catalytic effect as outlined by Norcini et al. in their criteria for good assessment[15].

The administrative burden of WBAs is a commonly cited barrier to implementation and acceptance[40,130]. Both feasibility and acceptability are critical components of good assessment, particularly formative assessment[15]. We worked to make the tool triggering and completion process as unobtrusive as possible to optimize feasibility and acceptability. Both residents and staff agreed the process was straightforward, easy, and that the tool itself was well designed and clear. However, residents did suggest having a more automated triggering process would help, while recognizing that this would require more advanced software or app functions. Because this tool was designed to provide feedback based on what residents did on-call, we believed it was also important for it to not get in the way of independent decision-making by residents. A desire for independence has been found to impact decisions whether to seek clinical support by trainees[49]. Our results did not find a negative impact of the tool on independence.

Another challenge in WBA implementation is a lack of understanding of the purpose of the assessment[38]. A critical principle of CBD is the focus on low-stakes formative assessments for learning instead of relying on summative assessments of learning[17,18]. However, medical trainees still often interpret formative assessments as having a summative intent, which negatively affects feedback-seeking and acceptance[45,46]. There were no instances of confusion with regards to the formative intent of our tool. This was a reassuring finding as residents seemed to feel comfortable receiving feedback through the tool. If the tool is to be applied in the future with more of a summative intent, perceived utility may decrease or there may be a trend towards a resident triggering an assessment only in the case of a more straightforward, less difficult case, to avoid negative feedback. This possibility was brought up by two interviewees and the triggering bias to avoid negative feedback is something that has been found to affect WBA uptake in other studies[38,127]. Our focus on creating a tool with a formative intent and making it clear to residents that it would not affect academic standing, seemed to help its acceptance.

An important consideration when designing this tool was that most supervision overnight is indirect. Without direct observation, entrustment decisions must rely on evidence of competence garnered from other sources, whether it be indirect interactions with

residents (e.g., on the phone), patients, the electronic medical record, post-operative visits, etc. Staff interviewees in our study reported they could provide better feedback using the tool if there was more interaction with the resident on-call for that specific case. Often this interaction was in the form of phone calls or text messaging. Although in-the-moment WBAs informed by direct observation form the cornerstone of a competency-based program[18,22,58], we know that direct observation occurs infrequently[63,65,131]. Furthermore, as the adoption of CBME continues, supervisors will need to engage in more frequent trainee assessments, and it is unrealistic to expect supervisors to directly observe each trainee interaction. Our results suggest that if a supervisor is assessing a trainee, but is unable to directly observed performance, ensuring some amount of communication with the trainee, even in the form of a quick case discussion by telephone, is beneficial. More communication between a resident and their assessor increases the information available to base an assessment on. In future studies, we plan to further explore what information staff members might be using to assign a score to each item. It would also be important to specifically examine what may make on-call related feedback credible from a resident perspective.

4.5 Conclusions

This study examined the impact, feasibility, and acceptability of the on-call tool developed in Chapter 3 through qualitative analysis of semi-structured interviews with residents and staff. Implementation of the tool was reported to have increased the amount of feedback given to residents based on their performance on-call and improved the structure of staff feedback. Staff were better able to complete the form and provide constructive feedback when they had more direct interaction with residents on-call. The tool triggering and completion process was considered simple and straightforward by both residents and staff and no negative consequences arose because of tool implementation.

A limitation of this study was potential selection bias relating to the interview participants. An inclusion criterion was that a participant must have completed or triggered at least one assessment. One resident commented how they did not send an assessment to a particular staff for feedback because they knew the assessment would not

be completed or would not contain much in terms of constructive feedback. This hesitancy to trigger assessments is a phenomenon described in previous medical education literature[98,99]. Therefore, only interviewing staff members who have completed at least one assessment, may have excluded those staff members who are less invested in the feedback-giving process. If there was an expectation in the future that this tool be completed by all staff members, more negative impressions of the tool could arise. Implementing a mandatory tool completion process may reveal negative consequences or other limitations which would have to be addressed. For any tool, there will be variable engagement from a staff and resident perspective.

Another possible limitation was that interview participants, particularly resident participants, may have been hesitant to share negative experiences with the interviewer (EM), who is a co-resident colleague. To limit this, as part of the consent process, anonymity of responses and the fact opinions would not affect academic standing, were emphasized. Having a co-resident colleague conduct interviews may have been better than a staff interviewer given the potential for a power differential[132,133]. An anonymous survey sent to pilot participants to gather impressions on the tool could be useful to corroborate some of the results of this qualitative study.

Chapter 5

5 Summary of Validity Evidence and Conclusions

5.1 Introduction

This final chapter is intended to summarize the validity evidence collected within this thesis. All chapters contributed validity evidence. We used Messick's framework to present our evidence in the categories of content evidence, response process evidence, internal structure evidence, relations to other variables evidence and consequences evidence[82]. The amount of evidence within each category varies, however, this is typical as validity should be thought of as existing on a spectrum and collection of evidence is an ongoing process[90]. As such, in this chapter we will also identify directions for future validity research and describe some overarching limitations within this thesis.

5.2 Validity Evidence

5.2.1 Content Evidence

Content evidence comes from ensuring the content within an assessment reflects the assessment's construct of interest[82,83]. The construct of our tool was surgical resident competence on-call. Time spent on-call makes up a significant portion of residency training and is considered instrumental to the development of clinical independence, however, there is minimal literature on what competence on-call entails[50]. A 2020 study of internal medicine residents sought to explore the experience of senior medical residents on-call overnight and found they must effectively perform in many domains including communication (with other health-care providers, nursing staff, and junior trainees), supervision of junior trainees, delegation of tasks, organization, and documentation[56]. Brady et al. designed the Orthopedic Intern Skills Assessment to simulate eleven clinical skills that an orthopedic resident would be required to do on-call in an attempt to assess whether interns were ready to transition to independent call without direct supervision[134]. All but one of these skills were procedural in nature. Another study team designed an On-Call Assessment Tool (OCAT) to evaluate what they considered three critical aspects of on-call performance: patient care, timeliness, and

sense of urgency[57]. These studies exhibit the breadth of required skills on-call. Ultimately, the success of a competency-based program relies on having assessments that reflect true clinical practice. So, understanding the important aspects of clinical practice on-call is crucial to design assessments specific to the on-call setting[23,135].

We used the consensus group methodology of NGT to harness the insights of surgical education experts to create tool content reflecting our construct of interest. NGT has been used in the past to develop items for assessment tools[35,72]. We believe our consensus group members were well equipped to discuss resident competence on-call. Two staff surgeon members were active program directors at the time of meeting, another two were former program directors, and the fifth staff surgeon was the CBME lead for the Division of PRS. The last group member was a senior surgical resident. The members represented three surgical specialties and the staff had an average academic hospital clinical experience of 17 years. The recommended steps for conducting a consensus group were followed, providing further credibility to the content development process[69]. The consensus group generated a list of ten items they agreed were essential components of surgical resident competence on-call.

The items from the consensus group can easily be linked to the CanMEDs roles used extensively in medical education literature (Table 18) (see Appendix 11 for CanMEDs role definitions).

Table 18. Linking CanMEDs roles to consensus group items

CanMEDs Role	Related Consensus Group Items
Medical Expert	Time management on call, recognition of urgency, appropriate use of backup, appropriate follow-up plan, overall management plan, clinical-judgement and/or decisions for treatment
Leader	Time management on call, recognition of urgency of presentation and appropriate follow-up plan

Collaborator	Handover process, documentation of encounter, and appropriate use of backup/knowing own limitations
Communicator	Communication with patient, development of rapport with patient
Professional	Development of rapport with patient
Scholar	None
Health Advocate	None

In the national survey from Chapter 2, PRS PDs were asked which CanMEDs roles they considered important to assess on-call. All agreed that medical expert, communicator, and leader were important. Four of five PDs agreed collaborator and professional were important to assess, while only two thought health advocate would be important, and only one selected scholar. In general, more consensus group items link to the CanMEDs roles deemed more important by program directors in the national survey (medical expert, communicator, and leader), further signifying their importance as aspects of on-call competence. There were no consensus group items that we believe link well to the health advocate or scholar roles. These two roles are likely better assessed in contexts other than the on-call period.

Two points of interest came up when defining the construct of interest. One was whether patient satisfaction was an essential part of competence on-call and the other was whether technical skills and clinical outcomes were important to assess.

Patient Satisfaction Content

All five PDs agreed the CanMEDs role of communicator was important to assess, however, only 40% of PDs specifically deemed patient satisfaction with overall encounter and patient satisfaction with communication as important to assess (Table 4). The CanMEDs role of communicator is actually specific to communication with patients, while the collaborator role relates more to communication with the medical team, so this discrepant finding from the survey may relate more to PDs being unclear as to the

CanMEDs role descriptions[12]. Clarification of the definitions and functions of each CanMEDs role within the survey may have resolved this discrepancy. In the national survey, residents noted a relative disinterest in receiving feedback on patient satisfaction with the overall on-call encounter (30% selecting they would want feedback in this area) and patient satisfaction with communication (14% selecting they would want feedback). Ultimately, two patient feedback-related items were included in the final consensus group list. There is a potential discordance between what residents want feedback on based on the national survey and what the consensus group members consider important to assess on-call.

Among medical educators there is agreement that patient feedback is important for resident learning and development[59,136,137]. Multi-source feedback, where feedback is obtained not just from supervisors, but from other medical team members and patients has been shown to benefit skills like communication and professionalism[137–140]. The low ranking of importance given to patient-related feedback in the national survey might reflect that while residents value this type of feedback, they may be skeptical of its accuracy and utility for educational improvement[141]. One resident interviewed in Chapter 4 reflected positively on the inclusion of patient feedback as a means of improving their communication. A qualitative study examining pediatric resident perspectives on patient and family feedback found residents thought they provided an important perspective on communication and interpersonal skills compared to feedback from other sources[142]. Based on our consensus group results, patient communication and professionalism are important aspects of on-call competence, however it will be important going forward to strike a balance between what residents want feedback on and what staff physicians think residents should get feedback on. Finding ways to improve the accuracy and credibility of patient feedback might improve its reception, potentially by having it mediated by a faculty coach or advisor[141]. The items included in our assessment could also be altered to meet the specific needs for a different program, specialty, or resident group.

Technical and Outcome-Related Content

Residents in the national survey valued feedback on clinical outcomes and technical-related decisions, but these aspects of care did not achieve consensus during the consensus group meeting. Consensus group members had a lengthy discussion about the items surgical adjuncts used, technical decision-making and procedural/clinical outcomes. Concern was raised about assigning a score for technical performance based on only seeing the result of, for example, a laceration repair or placement of a chest tube. It was thought the result may not always be reflective of the process that took place during the procedure. Gathering information from a procedure note was not thought to reliably reflect the true path taken either. For clinical outcomes, concerns were raised they would not always be assessable given the timeframe needed for some outcomes to occur. Within the interviews we held, some residents thought the tool was most applicable to procedurally based consults, and the tool had value in that it could provide feedback on outcomes in some situations. The existing literature suggests technical-related skills and decision-making benefits from in-the-moment teaching and coaching[26,143]. This is not to say that technical skills cannot be learned on-call, in fact, we believe that the on-call period is an excellent time to develop technical skills. Rather, assessment of technical skills may be better in other environments, like the operating room or in a simulation setting[26,29]. There is an obvious interest from residents for improved feedback on clinical outcomes and technical skills on-call, but the challenge is to determine the best way to achieve this. As it stands now, residents mostly hear about clinical outcomes through review of medical record reports or by reaching out to staff. Feedback on-call regarding technical skills should happen in the case of a senior resident supervising a junior or if a case goes to the operating room where staff will be present.

Additional Components of Competence On-Call

An important consideration is whether items generated in the working group encompass the extent of what surgical resident competence on-call entails. As suggested by Burm et al., inter-disciplinary communication with nursing staff, emergency physicians, outside hospital physicians, etc. is another critical aspect of effective performance for residents on-call[56]. While our tool looked to capture communication with patients and in the form of handover, communication with other team members was not captured. A limitation to

our method of defining the intended construct might have been the question we posed to consensus group members: "*what are the important aspects to include on a tool assessing surgical resident competence on-call?".* The addition of "…*include on a tool*…" within the question may have led to an emphasis of the competence aspects that are more feasible to assess (with a tool), rather than simply outlining the entire spectrum of competence on-call. As already discussed, technical ability is another aspect that working group members did not think was important to assess on-call and determined was better assessed through direct observation in other settings, however the ability of a surgical resident to effectively perform a procedure overnight would still contribute to competence on-call in a broader sense.

One of our thesis objectives was "to identify key elements of surgical resident competence on-call using consensus group methodology". We believe we identified many key elements, particularly the ones that working group members thought were the most important to assess on-call, but not all key elements. Without a doubt, the items in our tool do not capture the entirety of surgical resident competence on-call and this will have to be considered going forward.

To see whether the tool items reflect on-call competence in other surgical specialties we should explore the opinions of a larger sample of surgeons from other specialties and consider piloting the tool in other programs. Further revision of the tool based on emerging evidence in the literature regarding on-call competence must be done as well.

5.2.2 Response Evidence

Response process evidence examines "how well rater or examinee actions (responses) align with the intended construct"[91]. Part of the response process involves a rater's interpretation of the scoring system. For our tool, we created a novel construct-aligned scale, which provides evaluators with a standardized way of scoring based on the way day-to-day entrustment decisions are made, instead of relying on a traditional abstract scales (e.g., 1 = poor, 5 = excellent)[23]. The use of construct-aligned scales has been shown to result in more reliable and discriminating ratings compared to traditional scales[144] and they have been used widely in other medical education assessment

tools[29,35,106,144,145]. All staff members we interviewed in Chapter 4 commented on how they could appreciate the differences between the various scale levels. We did not specifically ask how their scoring decisions were made. Looking at the quantitative pilot results from Chapter 3, five of nine items had a minimum score given of two, which means staff were prepared to say the resident required in-person supervision to safely complete the task. All items had a negative skew (a bias towards higher ratings), which was to be expected given the level of training of the residents being assessed. The expectation for more senior PGY2 residents is that they can at least complete all necessary tasks on their own, but still may need higher levels of indirect supervision, which matches with the range of scores given. Including early PGY2 residents or PGY1 residents would almost certainly result in more scores from the lower end of the construct-aligned scale.

The construct of construct-aligned scales in medical education is most often some type of entrustment[146]. This was the case for our scale, as EM and AG reflected on the natural progression of entrustment on-call. Entrustment of residents on-call differs from daytime, as staff are not in hospital, and the onus is primarily on the resident to seek out supervision. Because staff are not in hospital, most supervision is indirect, which is what we found reported in the national survey. The scale we developed contained three tiers of indirect supervision on a spectrum of requiring some direction from staff to no direction from staff.

Day-to-day entrustment decisions on-call are made retrospectively, which was reflected in the wording of our scale. Retrospective scales reflect prior performance (e.g., how much supervision was required or provided to the trainee), whereas prospective scales require a rater to think ahead and make decisions about trust, which is a significant responsibility[147]. Either type can be used depending on the purpose of the tool. Again, from our qualitative interviews, staff agreed that the scale we designed accurately reflected the progression of entrustment on-call for residents. The response process for raters was hopefully optimized by using this intuitive entrustment construct-aligned scale.

Another way to optimize response process suggested in the literature is rater training. The act of conducting rater training alone does not equate to response process evidence, instead a study has to examine the impact of rater training[83]. The training given to raters in our study was limited. Orientation to the study and tool was done at the beginning of a scheduled academic meeting, however, the construct-aligned scale and the items were not specifically discussed. We also sent a follow-up email to staff members going over the study and tool basics again. Although intuitively rater training might make sense as a means of reducing score variance, this has not been born out in the medical education literature[148]. Rater training for the mini-CEX, probably the most studied WBA, did not find a significant improvement in inter-rater reliability or scoring accuracy[149]. The transition to using more construct-aligned scales may partly explain the lack of benefit seen from rater training.

Quality control of completed tools and scores is also part of the response process evidence. If data entry or reporting is inaccurate, ratings are not appropriately portrayed, impacting the response process. In our study, we had no instances of incomplete data within the completed forms. We did not, however, have a mechanism to ensure the number of forms returned to the study team for analysis equaled the number of forms completed by assessors. This would not be a problem if all assessments were completed online, but because we included the option to complete the assessment on paper or through an emailed pdf, completed tools could have not been returned to the study team. With multiple formats being available, it also raises the question as to whether assessors score differently depending on the format they are using, an area of potential research.

Going forward, we could complete think-aloud scoring with raters, where raters explain their decision-making behind items scores in-the-moment, to provide further response process evidence[150]. This would help with understanding what information raters are using to make their assessments, whether it be chart review, seeing the patient in person, observations made during the call period itself, etc.

Overall, the response process evidence for our study largely comes from the design and application of our novel construct-aligned scale. Both our qualitative and quantitative

results suggest the scale is easily interpreted, that scoring choices are logical, and that it accurately reflects the natural progression of entrustment on-call.

5.2.3 Internal Structure Evidence

Internal structure evidence examines the relations between items in a tool, their relation to the underlying construct, as well as results from generalizability studies and measures of reliability[77]. In this thesis, we collected item-level statistics and ran a generalizability study.

There were 63 occasions of tool completion across ten residents over a six-month period. Over six months there are approximately 180 call shifts, so for around 30% of those a tool was completed. This is slightly lower than the expected rate of completion, but there are multiple potential explanations for this. During some call shifts, residents may not have any consults at all, and even if a consult occurs, a resident may not always feel the need for feedback if it is something they are comfortable with. If a case went to the operating room overnight, hopefully feedback would be given in the moment, meaning the tool would be needed less to facilitate feedback. Residents did not receive any stipends or honoraria for completing the tools and part of the consent process was ensuring residents understood this was not mandatory and would not affect their academic standing, so this may have limited engagement. If the rate of tool completion is to be increased, the tool will likely have to be more formally implemented into the assessment curriculum.

Analyzing the items, the average score across all items was 4.03 ± 0.83. Five items had a minimum score of 2, with 3 being the minimum score on the other four items. This shows a clear tendency towards higher scores. Given we conducted this study with residents who are at the mid-PGY2 level and higher, it would have been concerning for any of the residents to get an item score of 1 (required complete takeover by a more senior physician). If the tool was used to assess PGY1 and early PGY2 residents as well, we would expect to see increased use of the lower end of the scale.

The overall reliability of the tool was 0.67. This reliability estimate includes the variance from all the measured facets. We found a generalization coefficient of 0.92 when looking at generalization between items (equivalent to internal consistency) and 0.28 when looking at generalization across occasions. For reference, the initial pilot of the O-SCORE had a reliability of 0.82 with 72 evaluations of 20 residents[29]. We believe the overall reliability of 0.67 for our tool is acceptable given this is pilot data obtained from a relatively small sample. Furthermore, much of the benefit of conducting a G-study derives from getting a more nuanced understanding of where the variance in scores comes from. From our G-study we can conclude that to increase reliability in our tool increasing the number of occasions would be more important compared to changing the number of items within the tool.

The finding of an internal consistency value of 0.92 is not particularly surprising as all tool items were considered representative of surgical resident competence on-call (based on the consensus group). In theory, the item scores should increase relatively uniformly as competence increases. A high internal consistency value such as this means assessment of our construct of interest could be done using fewer items. However, this was a consensus group developed list, and each item was considered an important element of on-call competence. So, for the purpose of formative feedback to residents, we believe all items should be retained.

A possible bias in our finding of high internal consistency and high average item scores could be the halo effect. This is where there is an inability to grade separate aspects of performance independently or where ratings are influenced by characteristics other than the targeted item or ability[118,119]. The halo effect is frequently seen in assessment data[151]. Adopting a program of assessments is a means of overcoming some of the possible effects of this bias.

Two other facets that would be useful to examine in the G-study are rater variance and case complexity variance. We were unable to measure rater variance and inter-rater reliability as there were no instances of multiple raters scoring a specific resident's performance based on a single patient interaction. This would be near impossible to

achieve as only one staff is on-call at a time. Score variance related to raters could be substantial for this tool. Case complexity could play a role in a different score being given for an individual resident across multiple occasions even if the rater is the same. We discussed including a box to indicate complexity of the presentation within the tool, however, decided this would be difficult to objectively determine.

Overall, the item analysis and G-study results provide internal structure evidence. The reliability of our tool was relatively low, but this is acceptable for a pilot study of a locally developed formative tool. Importantly, the results of the G-study suggest ways to improve tool reliability. As it stands, using our tool for summative assessment after only six assessments per resident, would be inappropriate given the reliability. This tool should be used in conjunction with other methods of assessment to understand resident competence more comprehensively.

5.2.4 Relations to Other Variables Evidence

Relations to other variables evidence looks at whether an assessment's scores align with external tests or measures examining a similar or related construct. Ideally, the external tests or measures will have their own validity evidence and be well-established[77]. The relationship can either be positive (strong association) or negative (lack of association).

There are no assessments or measures currently in place examining the construct of interest in our study, therefore we were not able to relate our results to another assessment looking at the same construct. We considered comparing against other assessment measures, like ITERs and EPAs. ITERs are the most common method of clinical evaluation in postgraduate medical education[152]. They involve collecting data and observing performance over an extended period and then creating an integrated assessment. Despite its widespread use, there are varying reports on the ability of ITERs to provide meaningful resident assessments[153,154]. The narrative comments found within ITERs have shown more promising validity evidence[155]. Regardless of the evidence for or against use of ITERs, given the format difference of ITERs, it would be difficult to realistically compare those scores with our tool scores. Relating to EPAs was another option; however, a major barrier to this was not all residents participating in the study

were in the CBD curriculum. EPA data was only available for PGY2 residents. Comparing the on-call assessment tool results to those from EPAs relating to a similar patient presentation could be an option in the future.

Given the lack of other assessments to use to relate to our scores, we decided to relate our scores to PGY-level of training. This has been reported in other studies as relations to other variables evidence[29,35,91,156]. We found our tool was able to differentiate between PGY2 and PGY4 or 5 residents, as well as between PGY3 and PGY 4 or PGY5 residents. There was no significant difference between average the scores of PGY2 and 3 residents or between the scores of PGY4 and 5 residents. The difference between more junior residents (PGY2 and 3) and more senior residents (PGY4 and 5) suggests the largest improvement in on-call competence occurs at the late PGY3 to mid PGY4 stage. This fits with responses from the semi-structured interviews. The highest burden of call occurs in PGY3, so it makes sense that the largest gains in on-call competence may occur around then, and this is when the tool was considered to be most beneficial or applicable by the residents we interviewed. In comparison, fourth- or fifth-year residents might be comfortable with the common on-call encounters, making the educational gains of a night shift smaller. The results from the G-study in Chapter 3 also show that the largest portion of the variance seen between scores seen is due to differences in PGY level. This is expected for a tool that can differentiate between PGY levels.

Relations to other variables evidence in this thesis comes from comparing tool scores to PGY level. We found a significant difference in scores with increasing PGY level. Going forward, tool scores should be compared to other measures like individual EPAs.

5.2.5 Consequences Evidence

Consequences evidence refers to the impact of the assessment and the assessment process, including beneficial or harmful effects. Despite often being under-examined, consequences evidence is an essential component of the validity argument[85,157]. We performed a reflexive thematic analysis of semi-structured interviews with pilot participants to gather evidence. This allowed us to collect stronger evidence compared to just using informal participant anecdotes[85]. We found both residents and staff members

considered the tool useful in improving the structure and quantity of feedback provided. To provide a reference point, we presented the impact of the tool in the context of the baseline feedback given to residents. We did not find any negative consequences resulting from tool use.

Residents found using the tool increased the amount of feedback they received. Staff found having a tool to fill out prompted them to give more feedback. As a result of the limited amount of baseline feedback, residents welcomed any improvement in the amount of feedback given. Most resident respondents from the national survey indicated they thought they would benefit from a more formal way of receiving feedback, and our tool appeared to make progress toward that goal. We did not elucidate the specific changes in resident performance or behaviour that may have resulted because of increased feedback.

Measuring the change in performance as a result of WBA implementation has been a long-standing challenge in the era of CBME. A 2012 systematic review found few quality studies showing an effect of feedback from WBAs on future performance in postgraduate medical education[158]. A more specific benefit of the tool in our case seemed to be that it facilitated the feedback process. Characteristics of effective feedback include being specific, timely, actionable and task-oriented and the baseline level of feedback residents described seemed to satisfy none of those features[42]. The tool served as an impetus for staff to pause and reflect on a resident's performance when an assessment was sent to them for completion, which could result in more actionable and specific feedback.

Identifying the particular tool aspects residents consider useful would be an important part of consequences evidence. Recent work has emphasized the importance of narrative comments[159–161]. In our study, 92% of tools had at least some text included in the narrative comments section. Exactly what was in these comments was not examined. While developing our tool we intentionally moved the narrative comment section to the middle of the assessment from the bottom in our original draft, to try and maximize the rate of completion. Including a section for patient feedback on communication and overall satisfaction was positively viewed by residents, however, this was only completed in 32% of cases. Again, the specific benefit of this part of the tool was not studied.

There were no reported anticipated or unanticipated negative consequences of the tool. We were particularly interested to see whether the tool had a negative impact on perceived autonomy of practice or whether there would be issues with acceptability and feasibility. A resident has increased autonomy and less supervision on-call compared to daytime clinical activities. Autonomy, anecdotally, helps build resident confidence, decision-making abilities and plays a role in preparation for independent future practice[50]. We did not find that our assessment tool adversely affected independence and autonomy on-call. Given the formative nature of the tool, the decision to trigger was made by the resident and so they retained control over whether feedback was given as a result.

The general perception by both residents and staff of the tool triggering and completion process was that it was feasible and acceptable. Concerns regarding administrative burden and time constraints as a result of WBA implementation is well documented in the literature[40,130]. In our study, some residents brought up the potential for increased assessment burden on staff members, however, none of the staff we interviewed found the process onerous. All agreed it was straightforward and easy to complete. A study looking at the feasibility of implementing a daily WBA for feedback on cataract surgery performance also found residents were more concerned about the time or administrative burden of the assessment on behalf of staff, compared to staff themselves[162].

Selection bias may have played a role in the lack of negative consequences identified by the interviews. An invitation to participate in the interviews was sent to all staff who completed at least one assessment, but there were multiple staff who did not complete a single assessment and may not have been sent any assessments. If residents knew they were unlikely to get feedback of value back from a specific staff member, they would be unlikely to send a tool out for completion. If our tool is included in the program of assessment within our division and staff are obligated to complete assessments, we may see more varied opinions on the consequences of tool use.

In summary, we found evidence of positive tool consequences in our study, and a lack of negative consequences. These are promising findings; however, ongoing examination of consequences should occur as the tool is used in the future, especially if it becomes more

of an established assessment used by our program or others. Further work must be done to try and characterize the positive impact and ascertain if there are sections of the tool which are more or less useful as a means of producing actionable and valuable feedback.

5.3 Limitations

Multiple limitations have already been discussed in previous chapters. We will touch on some broader limitations here.

Assessment of trainees is ideally done after direct observation[18,58,59], as direct observation is thought to improve the reliability and validity of clinical performance ratings and allow an assessment to include more actionable and specific feedback[61]. Direct observation rarely occurs during daytime hours, let alone on-call. This brings up a concern that assessments completed on residents based on what they do on-call would be less reliable, less valid and provide less meaningful feedback because of the lack of direct observation. This may be the case, however, our results show that residents still consider feedback given even after indirect observation to be beneficial, as do staff members. Not all WBAs will be appropriate for completion based on indirect observation. However, for our formative assessment looking at on-call competence, which was designed specifically with the understanding direct observation would not be possible, indirect observation appears to work. We acknowledge that summative decisions should not be made strictly based on these assessments, however they can contribute to the summative decision-making process. As mentioned before, exactly how assessors are completing the form based only on indirect observation is uncertain and needs to be examined.

A broader limitation exists with the reliance on indirect observation for assessment completion. Assessment completion primarily through indirect observation may be adequate to at least improve the assessment of competence on-call, as shown in this thesis. However, we believe using indirect observation alone is not sufficient to generate the most accurate assessment of competence on-call. For example, our tool asks staff to rate a resident's ability to triage. A staff rater may be able to gather some insight into this aspect through discussion with the resident and determining their own impression of the urgency of the presentation. To most accurately rate ability to triage though, staff would

probably have to listen in on the conversation between the resident and the emergency physician asking for a consult, for example. Also, as mentioned in section 5.2.1., the construct of surgical resident competence on-call is almost certainly broader than what was represented by items in our tool. If inter-disciplinary communication is also an important aspect of competence on-call, rating this this would require input from the health professionals a resident is communicating with, something indirect observation by a staff would be unable to capture. The most robust method of assessing competence on-call likely involves more multi-source feedback to increase direct observation and rely less on indirect observation.

The other question raised is whether our tool is generalizable. The working group we held did involve other surgical specialties, and we believe this did allow us to describe competence on-call more broadly. We did not, however, conduct the tool pilot in surgical specialties other than plastic surgery. We also did not survey residents or PDs from other specialties in Chapter 2 and did not interview residents or staff from other specialties in Chapter 4. Additionally, our program is quite small, having only 12 residents and 11 staff surgeons in total. We do think this tool is at least generalizable to other PRS programs based on our findings from the national survey as well as the fact the structure of call is similar across all PRS programs. The baseline culture of feedback and existing assessment patterns at our institution is likely similar to that in other PRS programs. The call structure for specialties like general surgery and orthopedics, for example, is different and the effect of differing call structures on the impact of our tool would be worthwhile to look at. As mentioned previously, part of the assessment design process is ongoing collection of validity evidence. Piloting our tool in plastic surgery was a realistic first step and expanding to other specialties can be done going forward.

5.4 Future Directions

Program of Assessments

Good assessment requires a programmatic approach[20]. Utilizing a combination of different assessments can alleviate the downsides of each individual assessment and allow the right assessment to be used at the right time for the right purpose[17]. The tool

developed in this thesis should not be used in isolation, but rather contribute as one component in a program of assessment. Program directors from other institutions and other specialties will have to decide whether our tool is applicable to their unique setting and appropriate for their own program needs. In our program, the tool can continue to be used as designed, and we also plan to work on ways to link the tool with the EPA system.

Future Format

We did see a variety of format preferences from both staff and residents, however from a long-term logistics and feasibility perspective, narrowing the number of ways to complete and trigger the assessment to just one electronic format is likely to happen. This would hopefully further simplify the documentation process, make feedback immediately available to residents, and allow for tracking over time. A way to automatically remind staff members about the fact they have an assessment to trigger, and potentially time this with when a patient is coming back to clinic, is much more feasible in an electronic assessment format versus paper format. Elentra (Elentra Corp, Kingston) is the software used in our program for completing and tracking of EPAs, so this could be a reasonable option to use.

Validity Evidence

There are many obvious areas for future research directions in terms of collecting more validity evidence. For content evidence, we could expand use of the tool to other institutions and other surgical specialties and elicit general feedback on the tool to ensure it represents the construct of interest. For response process evidence, we could carry out think-aloud studies and examine the effect of rater training. For internal structure evidence, we could examine what changes in our tool design or items might improve reliability. For relations to other variables evidence, we could look at the association of scores with relevant EPAs. For consequences evidence, we could see if there are parts of the tool that are more or less impactful in terms of providing useful feedback and whether the tool has quantifiable effects on performance or competence.

5.5 Conclusions

This thesis originated from an informal request from residents in our plastic surgery program for more feedback based on what they did on-call. We took this locally identified gap, and first confirmed its presence in other PRS programs across Canada through a national survey. We then used consensus group methodology to describe what competence on-call entails for surgical residents. We took the results of the consensus group and applied them using assessment development principles to create a novel formative assessment. We piloted this tool within our division and collected validity evidence using a modern validity framework to show that our tool does achieve its intended purpose of providing formative feedback to residents based on what they do on-call. Additionally, the consensus from staff members and residents was that the tool was feasible, acceptable and its use did not result in any negative consequences.

References

1. Potts JR. Shifting Sands of Surgical Education. *J Am Coll Surg*. 2018;227(2):151-162. doi:10.1016/j.jamcollsurg.2018.02.012

2. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting Paradigms: From Flexner to Competencies. *Acad Med*. 2002;77(5):361-367. doi:10.1097/00001888-200205000-00003

3. Tamblyn R. Outcomes in Medical Education: What is the Standard and Outcome of Care Delivered by our Graduates? *Adv Health Sci Educ*. 1999;4(1):9-25. doi:10.1023/A:1009893715930

4. HARDEN RM. AMEE Guide No. 14: Outcome-based education: Part 1-An introduction to outcome-based education. *Med Teach*. 1999;21(1):7-14. doi:10.1080/01421599979969

5. Leung WC. Competency based medical training: review. *BMJ*. 2002;325(7366):693-696.

6. Bell RH, Biester TW, Tabuenca A, et al. Operative experience of residents in US general surgery programs: a gap between expectation and experience. *Ann Surg*. 2009;249(5):719-724. doi:10.1097/SLA.0b013e3181a38e59

7. McGaghie WC, Miller GE, Sajid AW, Telder TV. Competency-based curriculum development on medical education: an introduction. *Public Health Pap*. 1978;(68):11-91.

8. Hamstra SJ. Keynote Address: The Focus on Competencies and Individual Learner Assessment as Emerging Themes in Medical Education Research. Kowalenko T, ed. *Acad Emerg Med*. 2012;19(12):1336-1343. doi:10.1111/acem.12021

9. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. *Med Teach*. 2010;32(8):638-645. doi:10.3109/0142159X.2010.501190

10. Sidhu RS, Grober ED, Musselman LJ, Reznick RK. Assessing competency in surgery: Where to begin? *Surgery*. 2004;135(1):6-20. doi:10.1016/S0039-6060(03)00154-5

11. Gordon M, Farnan J, Grafton-Clarke C, et al. Non-technical skills assessments in undergraduate medical education: A focused BEME systematic review: BEME Guide No. 54. *Med Teach*. 2019;41(7):732-745. doi:10.1080/0142159X.2018.1562166

12. Frank JR, Snell L, Sherbino J, Royal College of Physicians and Surgeons of Canada. *CanMEDS 2015: Physician Competency Framework*.; 2015.

13. Entrustability of professional activities and competency-based training - Ten Cate - 2005 - Medical Education - Wiley Online Library. Accessed December 20, 2021. https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2929.2005.02341.x

14. Edgar L, Holmboe E, McLean S, Hogan S, Hamstra S. The Milestones Guidebook. Published online 2020. https://www.acgme.org/portals/0/milestonesguidebook.pdf

15. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach*. 2011;33(3):206-214. doi:10.3109/0142159X.2011.551559

16. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med J Assoc Am Med Coll*. 1990;65(9 Suppl):S63-67. doi:10.1097/00001888-199009000-00045

17. Lockyer J, Carraccio C, Chan MK, et al. Core principles of assessment in competency-based medical education. *Med Teach*. 2017;39(6):609-616. doi:10.1080/0142159X.2017.1315082

18. Harris P, Bhanji F, Topps M, et al. Evolving concepts of assessment in a competency-based world. *Med Teach*. 2017;39(6):603-608. doi:10.1080/0142159X.2017.1315071

19. Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ*. 2005;39(1):98-106. doi:10.1111/j.1365-2929.2004.01972.x

20. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205-214. doi:10.3109/0142159X.2012.652239

21. Govaerts M, van der Vleuten CP. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47(12):1164-1174. doi:10.1111/medu.12289

22. Hatala R, Ginsburg S, Hauer KE, Gingerich A. Entrustment Ratings in Internal Medicine Training: Capturing Meaningful Supervision Decisions or Just Another Rating? *J Gen Intern Med*. 2019;34(5):740-743. doi:10.1007/s11606-019-04878-y

23. Gofton W, Dudek N, Barton G, Bhanji F. Workplace-Based Assessment Implementation Guide. Published online 2017. http://www.royalcollege.ca/rcsite/documents/cbd/wba-implementation-guide-tips-medical-teaching-practice-e.pdf

24. Harris KA, Nousiainen MT, Reznick R. Competency-based resident education—The Canadian perspective. *Surgery*. 2020;167(4):681-684. doi:10.1016/j.surg.2019.06.033

25. Bhatti NI, Cummings CW. Viewpoint: Competency in Surgical Residency Training: Defining and Raising the Bar: *Acad Med*. 2007;82(6):569-573. doi:10.1097/ACM.0b013e3180555bfb

26. Fritz T, Stachel N, Braun BJ. Evidence in surgical training – a review. *Innov Surg Sci*. 2019;4(1):7-13. doi:10.1515/iss-2018-0026

27. Vaidya A, Aydin A, Ridgley J, Raison N, Dasgupta P, Ahmed K. Current Status of Technical Skills Assessment Tools in Surgery: A Systematic Review. *J Surg Res*. 2020;246(k7b, 0376340):342-378. doi:10.1016/j.jss.2019.09.006

28. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278. doi:10.1046/j.1365-2168.1997.02502.x

29. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med J Assoc Am Med Coll*. 2012;87(10):1401-1407. doi:10.1097/ACM.0b013e3182677805

30. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107-113. doi:10.1016/j.amjsurg.2005.04.004

31. Doyle JD, Webber EM, Sidhu RS. A universal global rating scale for the evaluation of technical skills in the operating room. *Am J Surg*. 2007;193(5):551-555; discussion 555. doi:10.1016/j.amjsurg.2007.02.003

32. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *J Urol*. 2012;187(1):247-252. doi:10.1016/j.juro.2011.09.032

33. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. *Med Educ*. 2006;40(11):1098-1104. doi:10.1111/j.1365-2929.2006.02610.x

34. Hull L, Arora S, Kassab E, Kneebone R, Sevdalis N. Observational teamwork assessment for surgery: content validation and tool refinement. *J Am Coll Surg*. 2011;212(2):234-243.e1-5. doi:10.1016/j.jamcollsurg.2010.11.001

35. Rekman J, Hamstra SJ, Dudek N, Wood T, Seabrook C, Gofton W. A New Instrument for Assessing Resident Competence in Surgical Clinic: The Ottawa Clinic Assessment Tool. *J Surg Educ*. 2016;73(4):575-582. doi:10.1016/j.jsurg.2016.02.003

36. Parker SH, Flin R, McKinley A, Yule S. The Surgeons' Leadership Inventory (SLI): a taxonomy and rating system for surgeons' intraoperative leadership skills. *Am J Surg*. 2013;205(6):745-751. doi:10.1016/j.amjsurg.2012.02.020

37. Van Der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ*. 1996;1(1):41-67. doi:10.1007/BF00596229

38. Massie J, Ali JM. Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings. *Adv Health Sci Educ Theory Pract*. 2016;21(2):455-473. doi:10.1007/s10459-015-9614-0

39. Pereira EA, Dean BJ. British surgeons' experiences of mandatory online workplace-based assessment. *J R Soc Med*. 2009;102(7):287-293. doi:10.1258/jrsm.2009.080398

40. Driessen E, Scheele F. What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research. *Med Teach*. 2013;35(7):569-574. doi:10.3109/0142159X.2013.798403

41. Mann S, Truelove AH, Beesley T, Howden S, Egan R. Resident perceptions of Competency-Based Medical Education. *Can Med Educ J*. 2020;11(5):e31-e43. doi:10.36834/cmej.67958

42. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. *Med Educ*. 2019;53(1):76-85. doi:10.1111/medu.13645

43. Schut S, Maggio LA, Heeneman S, van Tartwijk J, van der Vleuten C, Driessen E. Where the rubber meets the road — An integrative review of programmatic assessment in health care professions education. *Perspect Med Educ*. 2021;10(1):6-13. doi:10.1007/s40037-020-00625-w

44. Beard J. Workplace-based assessment: the need for continued evaluation and refinement. *Surg J R Coll Surg Edinb Irel*. 2011;9 Suppl 1:S12-13. doi:10.1016/j.surge.2010.11.014

45. Harrison C, Wass V. The challenge of changing to an assessment for learning culture. *Med Educ*. 2016;50(7):704-706. doi:10.1111/medu.13058

46. Harrison CJ, Könings KD, Schuwirth LWT, Wass V, van der Vleuten CPM. Changing the culture of assessment: the dominance of the summative assessment paradigm. *BMC Med Educ*. 2017;17(1):73. doi:10.1186/s12909-017-0912-5

47. Kennedy TJT, Regehr G, Baker GR, Lingard LA. Progressive Independence in Clinical Training: A Tradition Worth Defending?: *Acad Med*. 2005;80(Supplement):S106-S111. doi:10.1097/00001888-200510001-00028

48. Haber LA, Lau CY, Sharpe BA, Arora VM, Farnan JM, Ranji SR. Effects of increased overnight supervision on resident education, decision-making, and autonomy. *J Hosp Med*. 2012;7(8):606-610. doi:10.1002/jhm.1959

49. Kennedy TJT, Regehr G, Baker GR, Lingard L. Preserving professional credibility: grounded theory study of medical trainees' requests for clinical support. *BMJ*. 2009;338(feb09 1):b128-b128. doi:10.1136/bmj.b128

50. Allen M, Gawad N, Park L, Raîche I. The Educational Role of Autonomy in Medical Training: A Scoping Review. *J Surg Res*. 2019;240:1-16. doi:10.1016/j.jss.2019.02.034

51. Halpern SD, Detsky AS. Graded Autonomy in Medical Education — Managing Things That Go Bump in the Night. *N Engl J Med*. 2014;370(12):1086-1089. doi:10.1056/NEJMp1315408

52. Morton JM, Baker CC, Farrell TM, et al. What do surgery residents do on their call nights? *Am J Surg*. 2004;188(3):225-229. doi:10.1016/j.amjsurg.2004.06.011

53. Drolet BC, Prsic A, Schmidt ST. Duty hours and home call: the experience of plastic surgery residents and fellows. *Plast Reconstr Surg*. 2014;133(5):1295-1302. doi:10.1097/PRS.0000000000000128

54. Walser E. Surgical Residency workload, perceptions and educational value: implications for competency- based medical education. *Electron Thesis Diss Repos*. Published online February 9, 2021. https://ir.lib.uwo.ca/etd/7743

55. Landmann A, Mahnken H, Antonoff MB, et al. Keeping Residents in the Dark: Do Night-Float Rotations Provide a Valuable Educational Experience? *J Surg Educ*. 2017;74(6):e67-e73. doi:10.1016/j.jsurg.2017.07.029

56. Burm S, Chahine S, Goldszmidt M. "Doing it Right" Overnight: a Multi-perspective Qualitative Study Exploring Senior Medical Resident Overnight Call. *J Gen Intern Med*. 2021;36(4):881-887. doi:10.1007/s11606-020-06284-1

57. Golnik KC, Lee AG, Carter K. Assessment of ophthalmology resident on-call performance. *Ophthalmology*. 2005;112(7):1242-1246. doi:10.1016/j.ophtha.2005.01.032

58. Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: The do's, don'ts and don't knows of direct observation of clinical skills in medical education. *Perspect Med Educ*. 2017;6(5):286-305. doi:10.1007/s40037-017-0376-7

59. LaDonna KA, Hatala R, Lingard L, Voyer S, Watling C. Staging a performance: learners' perceptions about direct observation during residency. *Med Educ*. 2017;51(5):498-510. doi:10.1111/medu.13232

60. St-Onge C, Chamberland M, Lévesque A, Varpio L. The role of the assessor: exploring the clinical supervisor's skill set. *Clin Teach*. 2014;11(3):209-213. doi:10.1111/tct.12126

61. Hasnain M, Connell KJ, Downing SM, Olthoff A, Yudkowsky R. Toward Meaningful Evaluation of Clinical Competence: The Role of Direct Observation in Clerkship Ratings: *Acad Med*. 2004;79(Supplement):S21-S24. doi:10.1097/00001888-200410001-00007

62. Burdick WP, Schoffstall J. Observation of emergency medicine residents at the bedside: how often does it happen? *Acad Emerg Med Off J Soc Acad Emerg Med*. 1995;2(10):909-913. doi:10.1111/j.1553-2712.1995.tb03108.x

63. Watling C, LaDonna KA, Lingard L, Voyer S, Hatala R. "Sometimes the work just needs to be done": socio-cultural influences on direct observation in medical training. *Med Educ*. 2016;50(10):1054-1064. doi:10.1111/medu.13062

64. Pulito AR, Donnelly MB, Plymale M, Mentzer RM. What do faculty observe of medical students' clinical performance? *Teach Learn Med*. 2006;18(2):99-104. doi:10.1207/s15328015tlm1802_2

65. Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Acad Med J Assoc Am Med Coll*. 2004;79(1):16-22. doi:10.1097/00001888-200401000-00006

66. Humphrey-Murto S, Varpio L, Wood TJ, et al. The Use of the Delphi and Other Consensus Group Methods in Medical Education Research: A Review. *Acad Med*. 2017;92(10):1491-1498. doi:10.1097/ACM.0000000000001812

67. Jones J, Hunter D. Qualitative Research: Consensus methods for medical and health services research. *BMJ*. 1995;311(7001):376-380. doi:10.1136/bmj.311.7001.376

68. Black N, Murphy M, Lamping D, et al. Consensus Development Methods: A Review of Best Practice in Creating Clinical Guidelines. *J Health Serv Res Policy*. 1999;4(4):236-248. doi:10.1177/135581969900400410

69. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and Nominal Group in medical education research. *Med Teach*. 2017;39(1):14-19. doi:10.1080/0142159X.2017.1245856

70. Manera K, Hanson CS, Gutman T, Tong A. Consensus Methods: Nominal Group Technique. In: Liamputtong P, ed. *Handbook of Research Methods in Health Social Sciences*. Springer Singapore; 2019:737-750. doi:10.1007/978-981-10-5251-4_100

71. Campbell SM, Cantrill JA. Consensus methods in prescribing research. *J Clin Pharm Ther*. Published online 2001:10.

72. Halman S, Dudek N, Wood T, et al. Direct Observation of Clinical Skills Feedback Scale: Development and Validity Evidence. *Teach Learn Med*. 2016;28(4):385-394. doi:10.1080/10401334.2016.1186552

73. Crenshaw K, Shewchuk RM, Qu H, et al. What should we include in a cultural competence curriculum? An emerging formative evaluation process to foster curriculum development. *Acad Med J Assoc Am Med Coll*. 2011;86(3):333-341. doi:10.1097/ACM.0b013e3182087314

74. Shortt SED, Guillemette JM, Duncan AM, Kirby F. Defining quality criteria for online continuing medical education modules using modified nominal group technique *. *J Contin Educ Health Prof*. 2010;30(4):246-250. doi:10.1002/chp.20089

75. Kelz RR, Sellers MM, Merkow R, Aggarwal R, Ko CY. Defining the Content for a Quality and Safety in Surgery Curriculum Using a Nominal Group Technique. *J Surg Educ*. 2019;76(3):795-801. doi:10.1016/j.jsurg.2018.10.005

76. Colón-Emeric CS, Bowlby L, Svetkey L. Establishing faculty needs and priorities for peer-mentoring groups using a nominal group technique. *Med Teach*. 2012;34(8):631-634. doi:10.3109/0142159X.2012.669084

77. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830-837. doi:10.1046/j.1365-2923.2003.01594.x

78. Yudkowsky R, ed. *Assessment in Health Professions Education*. 2nd edition. Routledge; 2020.

79. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. AERA; 1966.

80. Messick S. Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educ Res*. 1989;18(2):5-11. doi:10.3102/0013189X018002005

81. Kane M. Content-Related Validity Evidence in Test Development. In: *Handbook of Test Development*. Lawrence Erlbaum Associates Publishers; 2006:131-153.

82. Messick S. Standards of Validity and the Validity of Standards in Performance Asessment. *Educ Meas Issues Pract*. 2005;14(4):5-8. doi:10.1111/j.1745-3992.1995.tb00881.x

83. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul*. 2016;1(1):31. doi:10.1186/s41077-016-0033-y

84. Kinnear B, Kelleher M, May B, et al. Constructing a Validity Map for a Workplace-Based Assessment System: Cross-Walking Messick and Kane. *Acad Med*. 2021;96(7S):S64-S69. doi:10.1097/ACM.0000000000004112

85. Cook DA, Lineberry M. Consequences Validity Evidence: Evaluating the Impact of Educational Assessments. *Acad Med*. 2016;91(6):785-795. doi:10.1097/ACM.0000000000001114

86. Reliability: on the reproducibility of assessment data - Downing - 2004 - Medical Education - Wiley Online Library. Accessed March 12, 2022. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2929.2004.01932.x

87. Nunnally JC. An Overview of Psychological Measurement. In: Wolman BB, ed. *Clinical Diagnosis of Mental Disorders: A Handbook*. Springer US; 1978:97-146. doi:10.1007/978-1-4684-2490-4_4

88. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use, 5th Ed*. Oxford University Press; 2015:xiii, 399. doi:10.1093/med/9780199685219.001.0001

89. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49(6):560-575. doi:10.1111/medu.12678

90. Royal K. Four tenets of modern validity theory for medical education assessment and evaluation. *Adv Med Educ Pract*. 2017;Volume 8:567-570. doi:10.2147/AMEP.S139492

91. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ*. 2014;19(2):233-250. doi:10.1007/s10459-013-9458-4

92. Kogan JR, Holmboe ES, Hauer KE. Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review. *JAMA*. 2009;302(12):1316. doi:10.1001/jama.2009.1365

93. Goldstein MJ, Kim E, Widmann WD, Hardy MA. A 360 degrees evaluation of a night-float system for general surgery: a response to mandated work-hours reduction. *Curr Surg*. 2004;61(5):445-451. doi:10.1016/j.cursur.2004.03.013

94. Gordon WE, Gienapp AJ, Jones M, Michael LM, Klimo P. An Analysis of the On-Call Clinical Experience of a Junior Neurosurgical Resident. *Neurosurgery*. 2019;85(2):290-297. doi:10.1093/neuros/nyy248

95. McInnes CW, Vorstenbosch J, Chard R, Logsetty S, Buchel EW, Islur A. Canadian Plastic Surgery Resident Work Hour Restrictions: Practices and Perceptions of Residents and Program Directors. *Plast Surg*. 2018;26(1):11-17. doi:10.1177/2292550317749512

96. Sender Liberman A, Liberman M, Steinert Y, McLeod P, Meterissian S. Surgery residents and attending surgeons have different perceptions of feedback. *Med Teach*. 2005;27(5):470-472. doi:10.1080/01425590500129183

97. Bing-You RG. Why Medical Educators May Be Failing at Feedback. *JAMA*. 2009;302(12):1330. doi:10.1001/jama.2009.1393

98. Delva D, Sargeant J, Miller S, et al. Encouraging residents to seek feedback. *Med Teach*. 2013;35(12):e1625-e1631. doi:10.3109/0142159X.2013.806791

99. Reddy ST, Zegarek MH, Fromme HB, Ryan MS, Schumann SA, Harris IB. Barriers and Facilitators to Effective Feedback: A Qualitative Analysis of Data From Multispecialty Resident Focus Groups. *J Grad Med Educ*. 2015;7(2):214-219. doi:10.4300/JGME-D-14-00461.1

100. Thériault B, Marceau-Grimard M, Blais AS, Fradet V, Moore K, Cloutier J. Urology residents on call: Investigating the workload and relevance of calls. *Can Urol Assoc J J Assoc Urol Can*. 2018;12(2):E71-E75. doi:10.5489/cuaj.4333

101. Jackson JB, Huntington WP, Frick SL. Assessing the Value of Work Done by an Orthopedic Resident During Call. *J Grad Med Educ*. 2014;6(3):567-570. doi:10.4300/JGME-D-13-00370.1

102. Leafloor CW, Lochnan HA, Code C, et al. Time-motion studies of internal medicine residents' duty hours: a systematic review and meta-analysis. *Adv Med Educ Pract*. 2015;6:621-629. doi:10.2147/AMEP.S90568

103. Cunningham CT, Quan H, Hemmelgarn B, et al. Exploring physician specialist response rates to web-based surveys. *BMC Med Res Methodol*. 2015;15(1):32. doi:10.1186/s12874-015-0016-z

104. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med*. 2003;138(6):476-481. doi:10.7326/0003-4819-138-6-200303180-00012

105. Wilson AB, Choi JN, Torbeck LJ, Mellinger JD, Dunnington GL, Williams RG. Clinical Assessment and Management Examination--Outpatient (CAMEO): its validity and use in a surgical milestones paradigm. *J Surg Educ*. 2015;72(1):33-40. doi:10.1016/j.jsurg.2014.06.010

106. Warm EJ, Mathis BR, Held JD, et al. Entrustment and Mapping of Observable Practice Activities for Resident Assessment. *J Gen Intern Med*. 2014;29(8):1177-1182. doi:10.1007/s11606-014-2801-5

107. Kalet A, Zabar S, Szyld D, et al. A simulated "Night-onCall" to assess and address the readiness-for-internship of transitioning medical students. *Adv Simul*. 2017;2(1):13. doi:10.1186/s41077-017-0046-1

108. Whalen T, Wendel G. CHAPTER 6 NEW SUPERVISION STANDARDS: DISCUSSION AND JUSTIFICATION. Published online 2011:7.

109. Aylward M, Nixon J, Gladding S. An entrustable professional activity (EPA) for handoffs as a model for EPA assessment development. *Acad Med J Assoc Am Med Coll*. 2014;89(10):1335-1340. doi:10.1097/ACM.0000000000000317

110. Mink RB, Schwartz A, Herman BE, et al. Validity of Level of Supervision Scales for Assessing Pediatric Fellows on the Common Pediatric Subspecialty Entrustable Professional Activities. *Acad Med J Assoc Am Med Coll*. 2018;93(2):283-291. doi:10.1097/ACM.0000000000001820

111. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach*. 2012;34(11):960-992. doi:10.3109/0142159X.2012.703791

112. Makoul G, Krupat E, Chang CH. Measuring patient views of physician communication skills: development and testing of the Communication Assessment Tool. *Patient Educ Couns*. 2007;67(3):333-342. doi:10.1016/j.pec.2007.05.005

113. Agha RA, Fowler AJ, Sevdalis N. The role of non-technical skills in surgery. *Ann Med Surg*. 2015;4(4):422-427. doi:10.1016/j.amsu.2015.10.006

114. Dedy NJ. Teaching nontechnical skills in surgical residency: A systematic review of current approaches and outcomes. 2013;154(5):9.

115. Hamelin ND, Nikolis A, Armano J, Harris PG, Brutus JP. Evaluation of factors influencing confidence and trust in the patient-physician relationship: A survey of patient in a hand clinic. *Chir Main*. 2012;31(2):83-90. doi:10.1016/j.main.2012.01.005

116. Huda N, Faden L, Goldszmidt M. Entrustment of the on-call senior medical resident role: implications for patient safety and collective care. *BMC Med Educ*. 2017;17(1):121. doi:10.1186/s12909-017-0959-3

117. Transitioning towards senior medical resident: identification of the required competencies using consensus methodology - PubMed. Accessed July 19, 2022. https://pubmed.ncbi.nlm.nih.gov/30140348/

118. Thorndike EL. A constant error in psychological ratings. *J Appl Psychol*. 1920;4(1):25-29. doi:10.1037/h0071663

119. Cooper WH. Ubiquitous halo. *Psychol Bull*. 1981;90(2):218-244. doi:10.1037/0033-2909.90.2.218

120. MacEwan MJ, Dudek NL, Wood TJ, Gofton WT. Continued Validation of the O-SCORE (Ottawa Surgical Competency Operating Room Evaluation): Use in the Simulated Environment. *Teach Learn Med*. 2016;28(1):72-79. doi:10.1080/10401334.2015.1107483

121. Malhotra S, Hatala R, Courneya CA. Internal medicine residents' perceptions of the Mini-Clinical Evaluation Exercise. *Med Teach*. 2008;30(4):414-419. doi:10.1080/01421590801946962

122. Steiner I, Balsiger A, Goldszmidt M, Huwendiek S. Innovating Pediatric Emergency Care and Learning Through Interprofessional Briefing and Workplace-Based Assessment: A Qualitative Study. *Pediatr Emerg Care*. 2020;36(12):575-581. doi:10.1097/PEC.0000000000002218

123. Young JQ, Sugarman R, Schwartz J, O'Sullivan PS. Faculty and Resident Engagement With a Workplace-Based Assessment Tool: Use of Implementation Science to Explore Enablers and Barriers. *Acad Med J Assoc Am Med Coll*. 2020;95(12):1937-1944. doi:10.1097/ACM.0000000000003543

124. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77-101. doi:10.1191/1478088706qp063oa

125. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. *Med Teach*. 2020;42(8):846-854. doi:10.1080/0142159X.2020.1755030

126. Thematic Analysis: Striving to Meet the Trustworthiness Criteria - Lorelli S. Nowell, Jill M. Norris, Deborah E. White, Nancy J. Moules, 2017. Accessed November 7, 2022. https://journals.sagepub.com/doi/full/10.1177/1609406917733847

127. Gaunt A, Patel A, Rusius V, Royle TJ, Markham DH, Pawlikowska T. 'Playing the game': How do surgical trainees seek feedback using workplace-based assessment? *Med Educ*. 2017;51(9):953-962. doi:10.1111/medu.13380

128. Kornegay JG, Kraut A, Manthey D, et al. Feedback in Medical Education: A Critical Appraisal. Sherbino J, ed. *AEM Educ Train*. 2017;1(2):98-109. doi:10.1002/aet2.10024

129. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. *Perspect Med Educ*. 2015;4(6):284-299. doi:10.1007/s40037-015-0231-7

130. Bello RJ, Sarmiento S, Meyer ML, et al. Understanding Surgical Resident and Fellow Perspectives on Their Operative Performance Feedback Needs: A Qualitative Study. *J Surg Educ*. 2018;75(6):1498-1503. doi:10.1016/j.jsurg.2018.04.002

131. Burm S, Sebok-Syer SS, Van Koughnett JA, Watling CJ. Are we generating more assessments without added value? Surgical trainees' perceptions of and receptiveness to cross-specialty assessment. *Perspect Med Educ*. 2020;9(4):201-209. doi:10.1007/s40037-020-00594-0

132. Byrne E, Brugha R, Clarke E, Lavelle A, McGarvey A. Peer interviewing in medical education research: experiences and perceptions of student interviewers and interviewees. *BMC Res Notes*. 2015;8:513. doi:10.1186/s13104-015-1484-2

133. McGrath C, Palmgren PJ, Liljedahl M. Twelve tips for conducting qualitative research interviews. *Med Teach*. 2019;41(9):1002-1006. doi:10.1080/0142159X.2018.1497149

134. Brady JM, Smith D, Barronian T, et al. When Is an Orthopedic Intern Ready to Take Call? *J Surg Educ*. 2021;78(2):694-709. doi:10.1016/j.jsurg.2020.08.028

135. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach*. 2010;32(8):676-682. doi:10.3109/0142159X.2010.500704

136. Sargeant J, Bruce D, Campbell CM. Practicing physicians' needs for assessment and feedback as part of professional development. *J Contin Educ Health Prof*. 2013;33 Suppl 1:S54-62. doi:10.1002/chp.21202

137. Kogan JR, Holmboe E. Realizing the promise and importance of performance-based assessment. *Teach Learn Med*. 2013;25 Suppl 1:S68-74. doi:10.1080/10401334.2013.842912

138. Esteves A, McConnell M, Ferretti E, Garber A, Fung-Kee-Fung K. "When in Doubt, Ask the Patient": A Quantitative, Patient-Oriented Approach to Formative Assessment of CanMEDS Roles. *MedEdPORTAL J Teach Learn Resour*. 17:11169. doi:10.15766/mep_2374-8265.11169

139. Brinkman WB, Geraghty SR, Lanphear BP, et al. Effect of Multisource Feedback on Resident Communication Skills and Professionalism: A Randomized Controlled Trial. *Arch Pediatr Adolesc Med*. 2007;161(1):44. doi:10.1001/archpedi.161.1.44

140. Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad Med J Assoc Am Med Coll*. 2014;89(3):511-516. doi:10.1097/ACM.0000000000000147

141. Bogetz AL, Rassbach CE, Chan T, Blankenburg RL. Exploring the Educational Value of Patient Feedback: A Qualitative Analysis of Pediatric Residents' Perspectives. *Acad Pediatr*. 2017;17(1):4-8. doi:10.1016/j.acap.2016.10.020

142. Bogetz AL, Orlov N, Blankenburg R, Bhavaraju V, McQueen A, Rassbach C. How Residents Learn From Patient Feedback: A Multi-Institutional Qualitative Study of Pediatrics Residents' Perspectives. *J Grad Med Educ*. 2018;10(2):176-184. doi:10.4300/JGME-D-17-00447.1

143. Reznick RK, MacRae H. Teaching Surgical Skills — Changes in the Wind. *N Engl J Med*. 2006;355(25):2664-2669. doi:10.1056/NEJMra054785

144. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales:

Construct alignment improves workplace-based assessment scales. *Med Educ*. 2011;45(6):560-569. doi:10.1111/j.1365-2923.2010.03913.x

145. George BC, Teitelbaum EN, Meyerson SL, et al. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. *J Surg Educ*. 2014;71(6):e90-96. doi:10.1016/j.jsurg.2014.06.018

146. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med J Assoc Am Med Coll*. 2007;82(6):542-547. doi:10.1097/ACM.0b013e31805559c7

147. Ten Cate O, Schwartz A, Chen HC. Assessing Trainees and Making Entrustment Decisions: On the Nature and Use of Entrustment-Supervision Scales. *Acad Med J Assoc Am Med Coll*. 2020;95(11):1662-1669. doi:10.1097/ACM.0000000000003427

148. Vergis A, Leung C, Roberston R. Rater Training in Medical Education: A Scoping Review. *Cureus*. 2020;12(11):e11363. doi:10.7759/cureus.11363

149. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. 2009;24(1):74-79. doi:10.1007/s11606-008-0842-3

150. Padilla JL, Benítez I. Validity evidence based on response processes. *Psicothema*. 2014;(26.1):136-144. doi:10.7334/psicothema2013.259

151. Sherbino J, Norman G. On Rating Angels: The Halo Effect and Straight Line Scoring. *J Grad Med Educ*. 2017;9(6):721-723. doi:10.4300/JGME-D-17-00644.1

152. Chou S, Lockyer J, Cole G, McLaughlin K. Assessing postgraduate trainees in Canada: are we achieving diversity in methods? *Med Teach*. 2009;31(2):e58-63. doi:10.1080/01421590802512938

153. Watling CJ, Kenyon CF, Schulz V, Goldszmidt MA, Zibrowski E, Lingard L. An Exploration of Faculty Perspectives on the In-Training Evaluation of Residents. *Acad Med*. 2010;85(7):1157-1162. doi:10.1097/ACM.0b013e3181e19722

154. Kolars JC, McDonald FS, Subhiyah RG, Edson RS. Knowledge base evaluation of medicine residents on the gastroenterology service: Implications for competency assessments by faculty. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. 2003;1(1):64-68. doi:10.1053/jcgh.2003.50010

155. Hatala R, Sawatsky AP, Dudek N, Ginsburg S, Cook DA. Using In-Training Evaluation Report (ITER) Qualitative Comments to Assess Medical Students and Residents: A Systematic Review. *Acad Med J Assoc Am Med Coll*. 2017;92(6):868-879. doi:10.1097/ACM.0000000000001506

156. Jirativanont T, Raksamani K, Aroonpruksakul N, Apidechakul P, Suraseranivongse S. Validity evidence of non-technical skills assessment instruments in simulated anaesthesia crisis management. *Anaesth Intensive Care*. 2017;45(4):469-475.

157. Beckman TJ, Cook DA, Mandrekar JN. What is the Validity Evidence for Assessments of Clinical Teaching? *J Gen Intern Med*. 2005;20(12):1159-1164. doi:10.1111/j.1525-1497.2005.0258.x

158. Saedon H, Salleh S, Balakrishnan A, Imray CHE, Saedon M. The role of feedback in improving the effectiveness of workplace based assessments: a systematic review. *BMC Med Educ*. 2012;12(101088679):25. doi:10.1186/1472-6920-12-25

159. Ginsburg S, van der Vleuten CP, Eva KW, Lingard L. Cracking the code: residents' interpretations of written assessment comments. *Med Educ*. 2017;51(4):401-410. doi:10.1111/medu.13158

160. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ*. 2015;49(3):296-306. doi:10.1111/medu.12637

161. Ginsburg S, van der Vleuten CPM, Eva KW. The Hidden Value of Narrative Comments for Assessment: A Quantitative Reliability Analysis of Qualitative Data. *Acad Med J Assoc Am Med Coll*. 2017;92(11):1617-1621. doi:10.1097/ACM.0000000000001669

162. Nathoo NA, Sidhu R, Gingerich A. Educational Impact Drives Feasibility of Implementing Daily Assessment in the Workplace. *Teach Learn Med*. 2020;32(4):389-398. doi:10.1080/10401334.2020.1729162

Appendices

Appendix 1. REB approval for national survey

Western Research

Date: 22 July 2021

To: Dr. Aaron Grant

Project ID: 119109

Study Title: Assessing the current state of feedback for on-call encounters in plastic and reconstructive surgery

Application Type: HSREB Initial Application

Review Type: Delegated

Full Board Reporting Date: 08/Aug/2021

Date Approval Issued: 22/Jul/2021

REB Approval Expiry Date: 22/Jul/2022

Dear Dr. Aaron Grant

The Western University Health Science Research Ethics Board (HSREB) has reviewed and approved the above mentioned study as described in the WREM application form, as of the HSREB Initial Approval Date noted above. This research study is to be conducted by the investigator noted above. **All other required institutional approvals and mandated training must also be obtained prior to the conduct of the study**.

Documents Approved:

Document Name	Document Type	Document Date	Document Version
Redcap_Survey	Online Survey	21/Jun/2021	1
Redcap_Survey_Word	Online Survey	18/Jul/2021	1
Email Invitation_admin	Email Script	19/Jul/2021	1
OnCallFeedback Protocol_v3_clean	Protocol	18/Jul/2021	3
3 Week Follow-Up Email Invitation_v2	Email Script	20/Jul/2021	2
5 Week Follow-Up Email Invitation_v2	Email Script	20/Jul/2021	2
Email Invitation_v3_clean	Email Script	20/Jul/2021	3
Consent-LOI_v2_clean	Written Consent/Assent	20/Jul/2021	2

No deviations from, or changes to, the protocol or WREM application should be initiated without prior written approval of an appropriate amendment from Western HSREB , except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

REB members involved in the research project do not participate in the review, discussion or decision.

The Western University HSREB operates in compliance with, and is constituted in accordance with, the requirements of the TriCouncil Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2); the International Conference on Harmonisation Good Clinical Practice Consolidated Guideline (ICH GCP); Part C, Division 5 of the Food and Drug Regulations; Part 4 of the Natural Health Products Regulations; Part 3 of the Medical Devices Regulations and the provisions of the Ontario Personal Health Information Protection Act (PHIPA 2004) and its applicable regulations. The HSREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000940.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Karen Gopaul , Ethics Officer on behalf of Dr. Emma Duerden, HSREB Vice-Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Appendix 2. Email script for potential survey participants

Email Invitation

Dear colleagues,

You are being invited to participate in a study assessing the current state of feedback for on-call encounters in Plastic and Reconstructive Surgery programs across Canada. This study involves a 5-minute anonymous survey. All questions are completely voluntary. Participation is directed at program directors as well as PGY2 to PGY5 residents.

To complete the survey and review the Letter of Information please click here or copy and paste the link into your browser: https://redcap.lawsonresearch.ca/surveys/?s=7J8K7TJXNN

If you have any questions or concerns, please contact Eric Mitchell at [redacted] or [redacted] Please note that email is not a secure form of communication.

Thank you,

Eric Mitchell MD
Resident, Department of Plastic and Reconstructive Surgery, Western University

Aaron Grant MD MEd
Program Director, Department of Plastic and Reconstructive Surgery, Western University

Appendix 3. National survey questions

Survey

Please complete the survey below.

Thank you!

Are you a:	◯ Program Director ◯ Resident PGY 2 - 5
What year of residency training are you currently in?	◯ PGY2 ◯ PGY3 ◯ PGY4 ◯ PGY5
Approximately how many procedures do you perform on-call per month? (examples of procedures include soft tissue repair, wound debridement, tendon repair, fracture reduction and splinting, etc.)	◯ None ◯ 1-4 ◯ 5-9 ◯ 10-14 ◯ 15 or more
What level of supervision do you usually receive on-call, not including cases that go to the operating room?	◯ No supervision (no communication with on-call staff until morning handover) ◯ Indirect supervision (communication with on-call staff remotely by phone/text) ◯ Direct supervision (on-call staff present in person for at least part of call shift) ◯ Other (explain below)
Other	_____
What level of indirect supervision best describes what you usually receive on-call, not including cases that go to the operating room?	◯ I always discuss management plans with the on-call staff ◯ I often discuss management plans with the on-call staff ◯ I rarely discuss management plans with the on-call staff ◯ Other (explain below)
Other	_____
What is your level of satisfaction with the amount of feedback you receive from staff based on on-call encounters?	◯ Very unsatisfied ◯ Unsatisfied ◯ Neutral ◯ Satisfied ◯ Very satisfied
What is your level of satisfaction with the quality of feedback you receive from staff based on on-call encounters?	◯ Very unsatisfied ◯ Unsatisfied ◯ Neutral ◯ Satisfied ◯ Very satisfied
Does your program currently use any tools or forms to provide feedback to residents on-call?	◯ Yes (please explain below) ◯ No ◯ Unsure

Please explain which tools or forms are used	_____
Do you believe you would benefit as a resident from having a more formal method of receiving feedback?	◯ Yes ◯ No ◯ Unsure
How does your program currently assess resident on-call performance? (outside of operating room cases)	_____
Do you think there is room for more feedback to be given to residents based on on-call performance?	◯ Yes ◯ No ◯ Unsure
Does your program currently use a tool or form to provide feedback to residents based on on-call performance?	◯ Yes (if so, please describe what is used below) ◯ No ◯ Unsure
Please describe what tools or forms are currently being used	_____
Do you think residents and staff would benefit from having a tool or form to provide feedback to residents based on on-call performance?	◯ Yes ◯ No ◯ Unsure

Appendix 4. Open-ended responses from national survey

PD	Staff are not necessarily at the same site as residents making formal debriefing/evaluation difficult. A standard form may be helpful in structuring feedback but further evaluations on top of EPAs and O-scores may be onerous for residents
PD	I think that call is one of the few opportunities residents have to triage, manage, and communicate with minimal in-person supervision. This is very important in fostering independence and self-assessment. Any tool devised should not take away from that (i.e., I do not think call is the time to have witnessed encounters or more supervision).
Resident PGY2	I think it would be onerous to have feedback after every call shift (usually 9 a month as a junior). It would, however, be helpful if it were formally set up that at the end of the block each staff gave a bit of feedback on things done well, things to improve on.
Resident PGY2	Ideal time may be when patients return for follow up with the staff on call, could be discussed with the resident who saw them about what was/was not done well/outcomes/what can be done to improve management next time
Resident PGY3	On weekend calls I will see 5-7 consults or more and during week day evenings 1-2 usually with peripheral consults via the phone as well. If a tool were to be developed it should not be something that is a make work instrument for residents. It needs to have staff buy in or it will just add to the administrative burned and make me sad.
Resident PGY4	I appreciate when staff reward us for the work we do on call by taking time to teach or discuss a case. No feedback/teaching and a request for the demographic details to aid their billing leaves a sour taste
Resident PGY5	Some sort of formalized feedback or patient follow-up tool would be nice. At current time I informally follow my patients on call using the EMR or texting attendings but need to piece together how they are doing based on clinical notes and a more formal/complete information would be nice to refine my decision making

Resident PGY5	I just ask the staff directly if they agree or not with my management plan (for more difficult or instances that I am uncertain) and get direct feedback that way. Better than another form to fill out and faster time to feedback and putting it into action.

Appendix 5. REB approval for assessment data collection and interviews

Western Research

Date: 17 February 2022

To: Dr. Aaron Grant

Project ID: 120338

Study Title: On-call assessment tool validity in plastic surgery

Application Type: HSREB Initial Application

Review Type: Delegated

Meeting Date / Full Board Reporting Date: 15/Mar/2022

Date Approval Issued: 17/Feb/2022

REB Approval Expiry Date: 17/Feb/2023

Dear Dr. Aaron Grant

The Western University Health Science Research Ethics Board (HSREB) has reviewed and approved the above mentioned study as described in the WREM application form, as of the HSREB Initial Approval Date noted above. This research study is to be conducted by the investigator noted above. **All other required institutional approvals and mandated training must also be obtained prior to the conduct of the study**.

Documents Approved:

Document Name	Document Type	Document Date	Document Version
Interview Question Document_v1	Interview Guide	01/Jan/2022	1
Protocol_v3_clean	Protocol	12/Feb/2022	3
On Call Tool Data_v2	Other Data Collection Instruments	12/Feb/2022	2
Email-Invitation Assessments_v2_clean	Email Script	12/Feb/2022	2
Email-Invitation Interviews_v2_clean	Email Script	12/Feb/2022	2
Email-Reminder Assessments_v1	Email Script	12/Feb/2022	1
Email-Reminder Interviews_v1	Email Script	12/Feb/2022	1
Letter of Information and Consent for Assessment Data_v3_clean	Written Consent/Assent	16/Feb/2022	3
Letter of Information and Consent for Interviews_v4_clean	Written Consent/Assent	16/Feb/2022	4

Documents Acknowledged:

Document Name	Document Type	Document Date	Document Version
References_v1	References	01/Jan/2022	1
Budget_v1	Study budget	01/Jan/2022	1

No deviations from, or changes to, the protocol or WREM application should be initiated without prior written approval of an appropriate amendment from Western HSREB , except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

REB members involved in the research project do not participate in the review, discussion or decision.

The Western University HSREB operates in compliance with, and is constituted in accordance with, the requirements of the TriCouncil Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2); the International Conference on Harmonisation Good Clinical Practice Consolidated Guideline (ICH GCP); Part C,

Appendix 6. Preliminary London On-Call Assessment Tool

LOCAT (London On-call Assessment Tool)

Name of Resident: _____ Staff Surgeon On-call: _____

Name of Patient: _____ Date of Procedure: _____

MRN: _____ Procedure Performed: _____

Date of Clinic Visit: _____ _____

The purpose of this scale is to evaluate the trainee's performance of this procedure based on both (1) patient satisfaction and (2) expert evaluation of both management plan and outcomes.

Level of patient satisfaction:

Please ask the patient to rate their experience with the resident who performed the procedure using the visual analog scale. Ask the patient to focus on the resident's bedside manner, professionalism, communication etc.
If the patient is unsatisfied, please leave written feedback below.

0	1	2	3	4	5
Pt would trust resident to repeat procedure					Pt would not trust resident to repeat procedure

Quality of surgical outcome:

(for example, consider: skin closure, tendon repair strength, adequate bone resection, presence of complications like infection)

1	2	3	4	5
Poor surgical outcome	Satisfactory surgical outcome	Average surgical outcome	Good surgical outcome	Great surgical outcome

Quality of surgical adjuncts: (for example, consider: quality of splint, splinting position)

1	2	3	4	5
Poor outcome from surgical adjunct	Satisfactory outcome from surgical adjunct	Neutral outcome from surgical adjunct	Good outcome from surgical adjunct	Great outcome from surgical adjunct

Quality of post-procedural plan: (for example, consider: follow up timeline, CCAC set up)

1	2	3	4	5
Inappropriate post-op plan	Satisfactory post-op plan	Neutral post-op plan	Good post- op plan	Great post-op plan

Overall assessment of resident's management of this on-call procedure:

1	2	3	4	5
Needs definite improvement	Acceptable management with multiple areas for improvement	Satisfactory management with some room for improvement	Good management appropriate to level of training	Excellent management above level of training

Additional feedback: _____

Appendix 7. Literature review search protocol

Database: Ovid MEDLINE(R) ALL <2000 to current>

Search Strategy:

---

1 professional competence/ or clinical competence/ or (clinical$ and (skill$ or expertise$ or competen$)).mp. [SHOULD BE ADDEDD? Competency-Based Education/] (205530)

2 exp *Education, Medical, Graduate/ or exp "Internship and Residency"/ or *"Fellowships and Scholarships"/ (71914)

3 (resident$ or fellow$ or residenc$).ti. or (resident$ or fellow$ or residenc$).ab. /freq=2 (133074)

4 ((biomedical or clinical or medical or resident$ or fellow$ or residenc$) adj5 rotation$).tw,kf. (4477)

5 ((biomedical or clinical or medical) adj5 (resident$ or fellow$ or residenc$)).tw,kw,kf. (17698)

6 exp *Physicians/ or (doctor$ or surgeon$ or general pract$ or GP$1 or physician$).ti. or (doctor$ or surgeon$ or general pract$ or GP$1 or physician$).ab. /freq=2 (489816)

7 (graduate$ adj3 (student$ or intern$1 or trainee$ or resident$)).tw,kw,kf. (7935)

8 or/2-7 (646991)

9 1 and 8 (49499)

10 (perform$ adj5 (overnight$ or over-night$ or night$ or on-call or oncall or call)).tw,kf. (3060)

11 (perform$ adj5 (in-clinic$ or in-practice$)).mp. (4564)

12 ((on-call or oncall) and (service or duty or duties or hours or shift$1 or system)).tw,kw,kf. (1972)

13 (transition$ adj3 practice$).tw,kw,kf. (1564)

14 (Independent$ adj3 practice$).tw,kw,kf. or autonomy$.tw,kf. or (without adj5 superv$).tw,kf. (36610)

15 (entrust$ or superv$ or independent$).ti. or (entrust$ or superv$ or independent$).ab. /freq=2 (291149)

16 (perform$ adj5 (entrust$ or superv$ or autonom$)).tw,kf. (4659)

17 (multi$ adj5 feedback$).tw,kf. (2588)

18 ((technical$ or operative$ or surgical$ or task$ specific$) adj3 skill$).ti. or ((technical$ or operative$ or surgical$ or task$ specific$) adj3 skill$).ab. /freq=3 (2530)

19 ((judgement$ or autonomy$ or entrust$ or superv$ or independent$ or feedback$) and ((technical$ or operative$ or surgical$ or task$ specific$) adj3 skill$)).tw,kf. (2050)

20 or/10-19 (344620)

21 9 and 20 (5013)

22 *educational measurement/ or *self-evaluation programs/ or *test taking skills/ (19122)

23 (scale$1 or check-list$ or checklist$ or instrument$ or tool$1 or leaflet$).ti. or (scale$1 or check-list$ or checklist$ or instrument$ or tool$1 or leaflet$).ab. /freq=2 (650522)

24 23 and (exam$ or evaluat$ or assess$ or measure$).tw,kf. (425920)

25 (exam$ or evaluat$ or assess$ or measure$).ti. or (exam$ or evaluat$ or assess$ or measure$).ab. /freq=2 (5185693)

26 25 and (scale$1 or check-list$ or checklist$ or instrument$ or tool$1 or leaflet$).tw,kf. (780872)

27 22 or 24 or 26 (919134)

28 1 and 8 and 20 and 27 (1584)

29 limit 28 to english language (1558)

30 validation studies/ or exp "Surveys and Questionnaires"/st or st.fs. or validat$.tw,kf,kw. (1353001)

31 ((exam$ or evaluat$ or assess$ or measure$) and (scale$1 or check-list$ or checklist$ or instrument$ or tool$1 or leaflet$)).ti. (44261)

32 or/30-31 (1382148)

33 29 and 32 (855)

****

Appendix 8. Background document for consensus group

Consensus Group Background Document October 11th, 2021

Assessment in Surgery

As a starting point to stimulate your thoughts on the Nominal Question, here is some background information on competency and assessment tools used by surgical programs.

ACGME Core Competencies[10]	CanMEDs Roles[11]
Patient Care	Medical expert
Medical knowledge	Communicator
Practice-based learning and improvement	Collaborator
Interpersonal and communication skills	Professional
Professionalism	Leader
Systems-based practice	Scholar
	Health advocate

Competency in surgery can be broadly categorized into Technical Skills and Non-Technical Skills.[12–14] Non-Technical skills can be further divided into social skills (communication, teamwork, leadership), cognitive skills (decision making, situational awareness) and personal resource factors (ability to cope with stress and fatigue).[12]

Examples of Assessment Tools Used in Surgical Residency

- Ottawa Surgical Competency Operating Room Evaluation (O-SCORE)[2]
  - Assessment of pre-procedural plan, case preparation, knowledge of procedural steps, technical performance, visuospatial skills, post-procedure plan, efficiency and flow, communication
- Ottawa Clinical Assessment Tool (OCAT)[5]
  - Assessment of history, physical, case presentation, differential diagnosis, management plan, patient/family communication, documentation within clinic, collaboration, time management, technical skills, situational awareness
- On-Call Assessment Tool (OCAT)[15]
  - Assessment of history documentation, exam documentation, assessment and plan, consultation promptness, resident's perceived urgency rating
  - Of note, the development of this tool was specific to ophthalmology and not based on assessment development principles
- Multi-Source Feedback Questions[16]
  - Assessment of physician overall clinical performance, professionalism skills

There are a few other aspects that will be important to keep in mind during the meeting:

1. The tool will be completed in clinic at the post-encounter visit between the on-call staff and patient (i.e., not immediately post encounter)
2. The tool is intended to be generalizable and be applicable to multiple surgical specialties and potentially broader.
3. Consider aspects of competency that can be uniquely examined on-call and may not be captured as well in other environments (e.g., the operating room or clinic)

Thanks, and please reach out with any questions in advance of our meeting.

References

1. Gofton W, Dudek N, Barton G, Bhanji F. Workplace-Based Assessment Implementation Guide. Published online 2017. http://www.royalcollege.ca/rcsite/documents/cbd/wba-implementation-guide-tips-medical-teaching-practice-e.pdf

2. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med J Assoc Am Med Coll*. 2012;87(10):1401-1407. doi:10.1097/ACM.0b013e3182677805

3. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107-113. doi:10.1016/j.amjsurg.2005.04.004

4. Toprak A, Luhanga U, Jones S, Winthrop A, McEwen L. Validation of a novel intraoperative assessment tool: The Surgical Procedure Feedback Rubric. *Am J Surg*. 2016;211(2):369-376. doi:10.1016/j.amjsurg.2015.08.032

5. Rekman J, Hamstra SJ, Dudek N, Wood T, Seabrook C, Gofton W. A New Instrument for Assessing Resident Competence in Surgical Clinic: The Ottawa Clinic Assessment Tool. *J Surg Educ*. 2016;73(4):575-582. doi:10.1016/j.jsurg.2016.02.003

6. van der Vleuten CPM. Medical education research: a vibrant community of research and education practice. *Med Educ*. 2014;48(8):761-767. doi:10.1111/medu.12508

7. Shortt SED, Guillemette J-M, Duncan AM, Kirby F. Defining quality criteria for online continuing medical education modules using modified nominal group technique *. *J Contin Educ Health Prof*. 2010;30(4):246-250. doi:10.1002/chp.20089

8. Ito C, Ota K, Matsuda M. Educational content in nurse education in Japan: A Delphi study. *Nurs Ethics*. 2011;18(3):441-454. doi:10.1177/0969733010385530

9. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and Nominal Group in medical education research. *Med Teach*. 2017;39(1):14-19. doi:10.1080/0142159X.2017.1245856

10. Edgar L, Holmboe E, McLean S, Hogan S, Hamstra S. The Milestones Guidebook. Published online 2020. https://www.acgme.org/portals/0/milestonesguidebook.pdf

11. Frank JR, Snell L, Sherbino J, Royal College of Physicians and Surgeons of Canada. *CanMEDS 2015: Physician Competency Framework*.; 2015.

12. Aydin A, Fisher R, Khan MS, Dasgupta P, Ahmed K. Training, assessment and accreditation in surgery. *Postgrad Med J*. 2017;93(1102):441-448. doi:10.1136/postgradmedj-2016-134701

13. Fritz T, Stachel N, Braun BJ. Evidence in surgical training – a review. *Innov Surg Sci*. 2019;4(1):7-13. doi:10.1515/iss-2018-0026

14. Sidhu RS, Grober ED, Musselman LJ, Reznick RK. Assessing competency in surgery: Where to begin? *Surgery*. 2004;135(1):6-20. doi:10.1016/S0039-6060(03)00154-5

15. Golnik KC, Lee AG, Carter K. Assessment of ophthalmology resident on-call performance. *Ophthalmology*. 2005;112(7):1242-1246. doi:10.1016/j.ophtha.2005.01.032

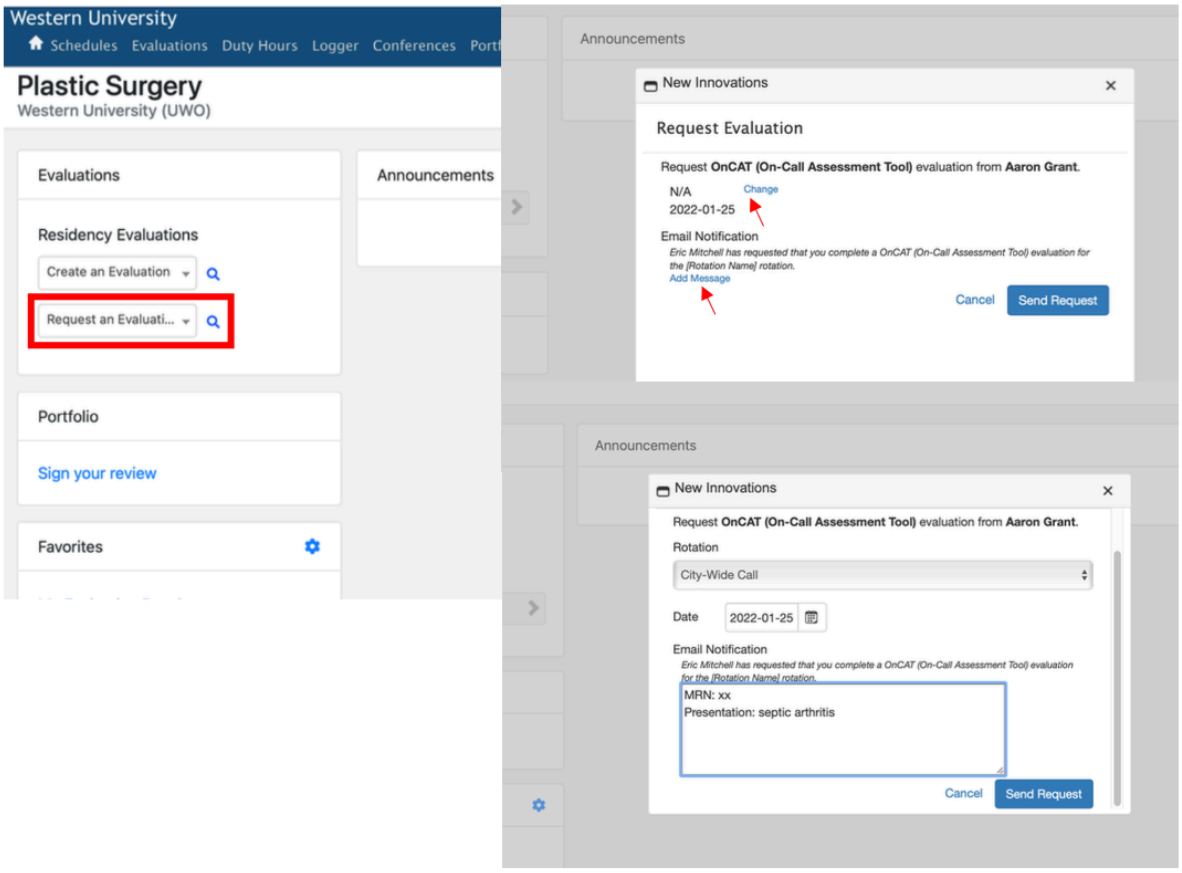
16. Murphy DJ, Bruce DA, Mercer SW, Eva KW. The reliability of workplace-based assessment in postgraduate medical education and training: a national evaluation in general practice in the United Kingdom. *Adv Health Sci Educ*. 2009;14(2):219-232. doi:10.1007/s10459-008-9104-8

Appendix 9. Email explanation of assessment triggering process

On-Call Assessment Tool Triggering

There are 3 methods of triggering completion of the on-call assessment tool:

1. Email completion of fillable pdf form
   - Either forward email from me or send a copy of the pdf to the staff member you were on call with
   - Please include **patient MRN** and **presentation** within the email as well as
2. Completion of printed copy in clinic
   - Currently available at UH, will provide copies for St. Joseph's and VH clinic spaces
   - Forward email from me or send request to staff by text/email, include **patient MRN** and **presentation**
3. New Innovations
   - There is now the option of triggering an assessment on New Innovations
   - Select "Request an Evaluation", and choose staff
   - Click "Change" to input date of patient assessment
   - Click "Add Message" and include **patient MRN** and **presentation** for identification



Appendix 10. Interview guide

Interview Questions

Introduction

Thank you for taking the time to meet with me today. The purpose of this meeting is to talk about your experience with the on-call assessment tool and investigate your thoughts on supervision and education on-call.

The interview should take between 30 and 60 minutes. I will be audio-recording the session because I do not want to miss any of your comments. All responses will be kept confidential. Your responses and the audio-recording will only be shared with the research team and a professional transcription service. We will ensure any identifying information is removed and the transcripts are anonymized.

Do you have any questions about any of this?

Are you willing to participate in the interview?

Resident Questions

1. What do you think the educational impact of being on-call is?
    a. Is the learning from an on-call shift different from during daytime hours?
2. What contributes to better learning on call?
3. What are potential barriers to learning on call?
4. What role does the staff supervisor overnight play in learning on-call?
    a. What are the attributes of an effective supervisor overnight?
5. What was your experience with the feedback tool?
6. Was the process of using the tool feasible?
7. Was there a difference in the feedback you received after implementation of the tool?

- Additional prompts
    - Change in quality of feedback?
    - Change in amount of feedback?
    - Did it help drive learning?
        - What part of it helped?
            i. Knowing the encounter would be assessed?
            ii. Getting the feedback itself?
            iii. Self-reflection afterwards?

8. Are there limitations to the feedback tool?
9. Any negative consequences of using the tool?
10. Do you think there's a difference in feedback given through this tool indirectly vs directly in person or over the phone?
11. Is there anything you would change about the tool or the process of using the tool?

Version 1
2022-01-01

Interview Questions

Staff Questions

Section 1

1. What do you think the educational impact of being on-call is?
    a. Do autonomy and independence play a role in learning on call?
        i. Is there a change from autonomy and independence during daytime hours?
2. What contributes to better learning on call?
3. What is a barrier to learning on call?
4. What role does the staff supervisor overnight play in learning on-call?
5. Does the educational impact of call or your learning objectives change over time in residency when moving from a more junior to senior resident?
6. What was your experience with the feedback tool?
7. Was using the tool feasible?
8. Did you have any specific concerns about any of the items or parts of the tool?
    a. E.g., clarity of items, not useful, not each to grade, etc.
9. What were your thoughts on the scale we used?
    a. E.g., was it easy to differentiate between a lower level on scale and upper level?
    b. Did you feel comfortable giving out low scores on the scale? Why or why not?
10. Do you think completing the tool made an impact on your feedback for residents?
11. How does this form compare to other EPAs you complete for residents?
12. Are there limitations to the feedback tool?
13. Any negative consequences of using the tool?
14. Is there anything else you would change about the tool or the process of using the tool?

Closing

Is there anything else you would like to add?

The next steps will be having the audio transcribed and to conduct an analysis of the transcription. I appreciate you taking the time to speak to me.

Appendix 11. CanMEDs role definitions

Communicator	As communicators, physicians form relationships with patients and their families that facilitate the gathering and sharing of essential information for effective health care.
Collaborator	As collaborators, physicians work effectively with other health care professionals to provide safe, high-quality, patient-centered care.
Leader	As leaders, physicians engage with others to contribute to a vision of a high-quality health care system and take responsibility for the delivery of excellent patient care through their activities as clinicians, administrators, scholars, or teachers.
Health Advocate	As Health Advocates, physicians contribute their expertise and influence as they work with communities or patient populations to improve health. They work with those they serve to determine and understand needs, speak on behalf of others when required, and support the mobilization of resources to effect change.
Professional	As Professionals, physicians are committed to the health and well-being of individual patients and society through ethical practice, high personal standards of behaviour, accountability to the profession and society, physician-led regulation, and maintenance of personal health.
Scholar	As Scholars, physicians demonstrate a lifelong commitment to excellence in practice through continuous learning and by teaching others, evaluating evidence, and contributing to scholarship.
Medical expert	Integration of the other 6 intrinsic CanMEDs roles

Curriculum Vitae – Eric Mitchell

Academic Background and Training

Plastic and Reconstructive Surgery Residency
Schulich School of Medicine & Dentistry, Western University
Current PGY3

Masters of Science (Surgery) - ongoing
Schulich School of Medicine & Dentistry, Western University
June 2021 – Dec 2022

Doctor of Medicine (MD)
Schulich School of Medicine & Dentistry, Western University
2016-2020

Bachelor of Science (Major – neuroscience, minor – economics)
McGill University
2012-2016

Publications

Haddara M, **Mitchell E**, Ferreira L, Gillis J, Suh N. The Evaluation of a FDP-to-Volar Plate Zone I Repair Versus Button Repair: An In-Vitro Biomechanics Study. *Accepted for publication, Journal of Hand Surgery.* 2022, Sep.

Silveira CRA, **Mitchell E**, Coleman K, Ruiz-Garcia R, Finger E. Changes in Motor Activity Level in Individuals with Frontotemporal Dementia. *In review.* 2022, Sep.

Mitchell E, Haddara M, Wu K.Y, Chambers S, Ferreira L, Gillis J. Comparison of Nerve Repair Technique in Upper Extremity Injuries: A Cadaveric Study. *Journal of Hand Surgery.* 2022, Feb.

Haddara M, **Mitchell E**, Ferreira L, Gillis J. The Effect of Flexor Digitorum Profundus Repair Position Relative to Camper's Chiasm on Tendon Biomechanics. *Journal of Hand Surgery.* 2021, Dec. (*co-first authors*)

ElHawary H, Salimi A, Alam P, Karir A, **Mitchell E**, Huynh M, Leveille C, Halyk L, St. Denis-Katz H, Iyer H, Padeanu S, Adibfar A, Valiquette C, Morzycki A, Janis J, Thibaudeau S. Gender Equality in Plastic Surgery Training: A Canadian Nationwide Cross-Sectional Analysis. *Plastic Surgery.* 2021, Nov.

Mitchell E, Tavares T, Palaniyappan L, Ahmed J, Finger E. Hoarding and Obsessive-Compulsive Behaviours in Frontotemporal Dementia: Clinical and Neuroanatomic Associations. *Cortex*. 2019, Oct.

Abstracts Presented (selected)

Development of an on-call assessment tool for use in plastic surgery. Poster presentation at the International Conference on Residency Education (ICRE), Oct 2022, Montreal, QC.

Development of an On-Call Assessment Tool for Surgical Specialties. Oral presentation at the 2022 Centre for Education Research and Innovation (CERI) Annual Research Symposium, Oct 2022, London, ON.

Biomechanical Evaluation of a Zone 1 FDP-to-Volar Plate Repair. Oral presentation at the Department of Surgery Research Day, June 2022, London, ON.

Biomechanical Evaluation of a Zone 1 FDP-to-Volar Plate Repair. Oral presentation at the Division of Plastic & Reconstructive Surgery Annual Resident Research Day, May 2022, London, ON.

Strength comparison of fibrin glue and suture constructs in upper extremity peripheral nerve coaptations: an *in-vitro* study. Virtual oral presentation at the American Association for Hand Surgery (AAHS) Annual Meeting, Jan 2022, Carlsbad, CA.

Strength comparison of fibrin glue and suture constructs in upper extremity peripheral nerve coaptations: an *in-vitro* study. Virtual oral presentation at the Department of Surgery Research Day, Nov 2021, London, ON.

The Effect of Flexor Digitorum Profundus Repair Position Relative to Camper's Chiasm on Work of Flexion and Tendon Loads. Oral presentation at the American Society of Plastic Surgeons (ASPS) Meeting, Oct 2021, Atlanta, GA.

The Effect of Flexor Digitorum Profundus Repair Position Relative to Camper's Chiasm on Work of Flexion and Tendon Loads. Oral presentation at the American Society for Surgery of the Hand (ASSH) Meeting, Sep 2021, San Francisco, CA.

The Effect of Flexor Digitorum Profundus Repair Position Relative to Camper's Chiasm on Work of Flexion and Tendon Loads. Virtual oral presentation at the Canadian Society of Plastic Surgeons (CSPS) Meeting, June 2021.

Hoarding and Obsessive-Compulsive Behaviours in Frontotemporal Dementia. Oral Presentation at the American Academy of Neurology (AAN) Annual Meeting 2019, Philadelphia, PA.

Naloxone kits at AEDs: A case-study of medical student driven advocacy in London, Ontario. Poster at the Canadian Conference on Medical Education (CCME) 2019 Meeting, Niagara Falls, ON.

Grants, Honours, and Awards

Resident Research Grant ($5000)
Competitive grant funding for project in patient-reported outcomes post-wrist trauma awarded by the Department of Surgery, Western University, Nov 2022.

David J. Hollomby Award
Awarded for top oral abstract at the Center for Education Research and Innovation (CERI) Symposium, Oct 2022, London, ON.

Canada Graduate Scholarship ($17,500)
Awarded by the Canadian Institute for Health Research (CIHR), May 2022.

McLachlin Resident Research Grant ($4438)
Competitive grant funding for project in surgical competency-based medical education awarded by the Department of Surgery, Western University, Dec 2021.

Best Presentation Award in Hand Session.
Awarded at the American Society of Plastic Surgeons (ASPS) meeting, Oct 2021, Atlanta, GA.

Top 5 Best Research Paper Award
Awarded at the American Society for Surgical of the Hand (ASSH) meeting, Sep 2021, San Francisco, CA.

Stanley Markell Dow Scholarship ($1100)
Awarded based on academic performance and financial need, Western University, 2019.

Summer Research Training Program Grant ($9000)
Funding awarded for participation in a research training program, Western University, 2017-2018.

Additional Ongoing Research Experience

Perilunate Injuries – Long Term Follow-Up of Outcomes and Disease Progression. St. Joseph's Health Care, London, ON. Supervisor: Dr Ruby Grewal.

Retrospective Evaluation of Fascicular-Targeted Supercharge End-to-Side (SETS) Anterior Interosseous–to–Ulnar Motor Nerve Transfer. St. Joseph's Health Care, London, ON. Supervisor: Dr Joshua Gillis.

Assessing the Educational Benefit of Time On-Call for Surgical Residents: A Qualitative Study. Western University, London, ON. Supervisor: Dr Aaron Grant.