Using the Ottawa Surgical Competency Operative Room Evaluation (O-SCORE) in a Canadian Plastic Surgery Program

by

Curtis Robert Budden

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

In

Measurement, Evaluation and Cognition

Department of Educational Psychology

University of Alberta

Abstract

Competency-based medical education has overtaken the traditional apprenticeship model of medical education. This shift toward competence-based training demands a higher number of assessments per trainee. Current assessment methods in Plastic Surgery have not been well studied. Numerous assessment tools in surgery exist but do not apply to the specialty of Plastic Surgery because of the nature of procedures being assessed. One tool which combines a focus on competency as well as possible applicability to Plastic Surgery is the Ottawa Surgical Competency Operative Room Evaluation (O-SCORE). This tool uses a unique entrustability scale which is construct aligned with competency-based training. The purpose of this study was to examine the validity evidence and reliability of assessments made using the O-SCORE in a Canadian Plastic Surgery training program.

Ten residents at the University of Alberta participated in this research and 41 evaluations were performed by supervising Plastic Surgeons. Residents were asked to be evaluated on three common Plastic Surgery procedures (a) breast reduction mammoplasty, (b) mandibular fracture open reduction internal fixation, and (c) and hand fracture fixation. The psychometric properties of the items were assessed using item-total correlation and Cronbach's alpha. Validity evidence was examined through the differentiation of trainees based on level of training. Reliability testing was performed in two ways. The first was a video analysis of technical skills by Plastic Surgeons. The second was a generalizability study. This G study was a mixed, two facet design. Facets were occasion and item. Occasion in this study was nested within resident as trainees were evaluated on different occasions.

The items on the O-SCORE were found to have high item-total correlations (range 0.74 to 0.88). Cronbach's alpha for technical items based on video analysis was 0.91. It was found that assessments using the O-SCORE significantly differentiated junior and senior trainees on all

items and overall score, $\Lambda = 0.08$, $F(4,36) = 2.77$, $p < 0.001$. The O-SCORE contains a dichotomous item assessing competency and this was found to be significantly different between junior and senior trainees, $\chi^2 = 29.73$, $\nu = 1$, $p < 0.001$. The dependability coefficient from the generalizability study was 0.91. A decision study was conducted and showed that a minimum of two occasions were necessary to obtain a reliability coefficient of greater than 0.85.

The O-SCORE and its entrustability-scaled anchors offer a construct-aligned assessment tool for operative skills. At the University of Alberta's Plastic Surgery residency program, it has been shown that this tool offers high internal consistency of technical items, high overall reliability with three occasions, and validity evidence. Assuring competence is the goal of CMBE and using the O-SCORE as part of an assessment armamentarium is a step closer to this objective.

Preface

This is an original work of Curtis Budden. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board. Project name, “Use of the O-SCORE for plastic surgery assessment”, Pro00057378, July 14, 2015.

Acknowledgements

This work would not have been possible without the participation of the University of Alberta Plastic Surgery residents and staff surgeons. I will always appreciate the willingness to continue to complete and have evaluations completed. Guiding me through this research process were my supervisors, Dr. Mark Gierl, Dr. Jonathan White and Dr. Jay Zhu.

I would like to thank Dr. Mark Gierl for his continued support for meaningful completion of this work. Always available to provide guidance, I am grateful for his expertise with this project.

Dr. Jonathan White has been very supportive of my work in this area. His expertise in clinical evaluation and familiarity with surgical video recording were a tremendous asset to this project.

Dr. Jay Zhu was instrumental in garnering support from the division of Plastic Surgery for this project. His ongoing support for this project was very much appreciated.

Next I would like to acknowledge the non-academic support provided by my sounding board and motivators Raphael, Allison, Quigs, Erica, Jessica, Sam, Courtney and Chris. Having the opportunity to discuss issues and struggles with you was a tremendous help. Thank you to my parents and brother Chad for their ongoing support and excitement for my endeavours.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

OSCORE	Ottawa Surgical Competency Operating Room Evaluation
CBME	Competency Based Medical Education
ACGME	American College Graduate Medical Education
PGME	Post Graduation Medical Education
PGY	Post-Graduate Year
EPA	Entrustable Professional Activity
CORE	Comprehensive operative resident evolution
SEM	Standard Error of the Mean

Competency-based medical education (CBME) demands valid and reliable assessment. The current apprenticeship model of medical training is being overtaken by competency-based training. The Royal College of Physicians and Surgeons of Canada has released a plan for "Competence by Design" which will be implemented across Canada in all specialties of training (Royal College of Physicians and Surgeons of Canada, 2014). This competency-based training model will place greater demand on training programs to perform regular assessment of resident performance. Based on a recent review of assessment tools, surgical educators should be aware that "there been considerable lag in the use of contemporary framework for conceptualizing validity, and that framework has not been adopted in surgical education" (Ghaderi et al., 2015). In plastic surgery there is a paucity of research on the reliability of assessment tools. There are few previously studied assessment tools which are applicable to Plastic Surgery. One of these potential tools is the Ottawa Surgical Operating Room Evaluation tool (Gofton, Dudek, Wood, Balaa, & Hamstra, 2012). This tool uses unique entrustability anchors to rate trainees. The purpose of this thesis research was to explore the reliability of the O-SCORE when used in Plastic Surgery resident training.

Overview of Competency-Based Training

The first published articles on competency-based education date back over 60 years (Frank, Snell et al., 2010; Grant, 1979). Prior to this time, medical training in North America took a dramatic turn in 1910 with the publication of the Flexner report (Flexner, 1910).This important work, commissioned by the Carnegie Foundation, resulted in numerous insights into medical education and recommendations for future training. Flexner proposed three factors which are necessary in medical education. These are (1) the creation of public awareness that all physicians should be trained in fundamentals of medical science, (2) standards for universities

and their relationship with medical schools, and (3) the profession's sense of duty toward high standards of practice (1910). The competency of medical school graduates was an issue in 1910 as much as it is today. Significant progress in medical education has been made over the past century. One quote from the Flexner report, however, illustrates a similar concern felt by postgraduate schools today: "the postgraduate school was thus originally an undergraduate repair shop. Its instruction was necessarily at once elementary and practical. There was no time to go back to fundamentals; it was too late to raise the question of preliminary educational competency" (Flexner, 1910). Postgraduate training programs today are seeking well-trained medical students for similar reasons as were present in 1910. There is much to learn and only a short period of time over which to learn it.

Since 1910, competency-based reform of medical education has been discussed in varying capacities (Frank et al., 2010). Although no one individual is accredited with developing CBME, many groups particularly in Canada, Netherlands, and the United States have made significant contributions to this area (C. L. Carraccio & Englander, 2013; Frank, Mungroo et al., 2010; ten Cate & Scheele, 2007a). Some of the earlier publications on CBME describe implementation and evaluation of a competency-based model (Spady, 1977). Carraccio, Wolfsthal, Englander, Ferentz, and Martin (2002) presented a review of competency-based training where they sought to "understand the evolution of the educational paradigm" (p. 361). In this paper the authors highlight the differences between the traditional medical training model and a competency-based training program. Features of CBME which contrast the traditional medical model are that the assessment tools should be authentic, should be done through direct observation, and should be criterion referenced. The O-SCORE assessment tool used in this study meets all these criteria and will be revisited later in this section.

The Rationale for CBME

When exploring why there has been such a strong emphasis on CBME in the recent years, Frank et al described four emerging themes. These are (a) a focus on curricular outcomes, (b) emphasis on abilities, (c) a de-emphasis of time-based training, and (d) the promotion of learner centeredness. These themes do not describe the underlying push for CBME, which is a public demand for increasingly accountability for the self-governing medical profession (C. Carraccio, Wolfsthal, Englander, Ferentz, & Martin, 2002). An additional rationale for CBME was proposed by Leung (2002) that in order to maintain competitiveness in the global workforce, local leaders in medical training needed to deliver programs with a competency-based framework that was matched to countries leading the way for medical education reform. Despite this list of factors leading to CBME, the main motivating factor remains the demand from the public for assurance of competent physicians.

Definition of Competency

To define competency-based training took many years (Frank et al., 2010). It was summarized that the intention of CBME is to develop "a health professional who can practice medicine at a defined level of proficiency, in accord with local conditions, to meet local needs" (McGathie, Miller, Sajid, & Telder, 1978). One of the earliest definitions of competency-based education is "a data-based, adaptive, performance-oriented set of integrated processes that facilitate, measure, record and certify within the context of flexible time parameters the demonstration of known, explicit[l]y stated, and agreed upon learning outcomes that reflect successful functioning in life roles" (Spady, 1977). This definition was developed from research in school-based learning in the United States. Despite this early definition there was still much debate regarding how the medical community should define competency. This is demonstrated

by the fact that from the years 2000 to 2016 there are approximately ten publications per year regarding the definition of CBME (Frank et al., 2010). The most recent operational definition of competency-based training was developed through a thematic analysis of 137 publications on the definition of CBME. It can be defined as "an approach to preparing physicians for practice that is fundamentally oriented to graduate outcome abilities and organized around competencies derived from an analysis of societal and patient needs. It de-emphasizes time-based training and promises greater accountability, flexibility, and learner-centeredness" (Frank et al., 2010). In comparison to the earlier definition, this one incorporates societal needs and summarizes the current purpose of CBME.

CBME and Plastic Surgery

Plastic Surgery is a specialty concerned with function. Translated from the Greek *plasticos*, which means reconstruction. The daily activities of a Plastic Surgeon are variable as this is a specialty encompassing many different areas. These areas include craniofacial, hand, peripheral nerve, burn, breast, cosmetic and pediatric surgery.

The adequacy of Canadian Plastic Surgery training has recently been explored (Chivers et al., 2013; Ferron, Lemaine, Leblanc, Nikolis, & Brutus, 2010). These authors explored perceived readiness for practice and satisfaction with training. Overall, graduating trainees are satisfied with their training. Cosmetic surgery training in Canada remains as an area of weakness despite constituting a large proportion of the Royal College licensing examination. In one study it was found that less than 20 percent of residents graduating in Canada felt prepared to incorporate cosmetic surgery into their professional practice (Chivers et al., 2013). Given this paucity of preparedness, one could ask how the competency of key cosmetic procedures is assessed. Watching a surgeon perform cosmetic procedures is hardly adequate to assess whether a resident

is competent to perform these procedures independently. The development of competency in this area is lacking. In addition to cosmetic surgery, it was published that some non-technical competencies are less well taught to trainees (Kasten, Levi, Eng, & Schenarts, 2009). As Plastic Surgery residency shifts to a competency-based training model, it is important that all key objectives are taught well to ensure skills are transferred and used appropriately by resident trainees. The assessment methods used in Plastic Surgery are heterogeneous and not standardized across programs (Kasten et al., 2009). The most commonly used assessment methods are in service examinations, scheduled formal verbal feedback, and global assessment scores.

Focusing on Plastic Surgery training, Bancroft et al. (2008) published their interpretation of outcomes-based residency education. They suggested that competencies for trainees can be assessed using a combination of global rating scale, 360 degree evaluations, checklists, oral examination, multiple choice questions, portfolios, and patient surveys.

As competency-based training is being implemented, there are some assessment details which seem to be well established. One of these details is that objectives will be defined by milestones and entrustable professional activities (EPA) (McGrath, 2014; ten Cate, 2005). As program administrators develop assessment strategies, it is known that certain domains or constructs are difficult to assess. An example of this is the evaluation of professionalism. Dividing professionalism into multiple aspects and basing an evaluation on a checklist risks destroying validity evidence and the authenticity of the construct (Ginsburg, McIlroy, Oulanova, Eva, & Regehr, 2010). This example represents one of the main criticisms of CBME, which is the reductionism of medical training (Norman, Van der Vleuten, & De Graaff, 1991). Based on the trend of using milestones and EPAs to guide CBME, recommendations for an assessment model in a competency-based Plastic Surgery residency program were described (Knox,

Gilardino, Kasten, Warren, & Anastakis, 2014). The authors highlighted how an assessment matrix in breast reconstruction could be utilized to comply with competency-based training. It emphasized the objectification of the main Plastic Surgery competencies. This is currently the only published approach to evaluation of Canadian Plastic Surgery trainees in a competency-based model. In the Unites States, the American College of Graduate Medical Education (ACGME) has published updates on the milestone project with regards to Plastic Surgery (McGrath, 2014). The milestone project is the equivalent of the Royal College of Physicians and Surgeons of Canada's Competence by Design initiative. The original Plastic Surgery milestones were tested in 21 training programs in the United States. The milestones working group published six assessment tools which programs have the option of using should they choose (Plastic Surgery Milestone Working Group.,). These assessment tools consist of checklists and one modified Observed Structures Assessment of Technical Skills (OSATS) style assessment tool. A review of published literature failed to yield any studies examining the psychometric properties of these tools nor are there studies on the validity of the assessments. By developing a CBME model and spending countless hours on subdividing a specialty into required competencies, failing to ensure that the assessment methodology is rigorous seems to negate the prior work. The specifications of competence by design model for Plastic Surgery in Canada have yet to be released. Nonetheless, it is imperative that expert panels enlist the expertise of those with a working knowledge of assessment and the intricacies of that complex task.

Assessment in Medicine

Leaders in medical education strive for quality in assessment. Performance assessment has received much attention in the literature prior to the push for CBME. In the 1990s much debate was had as to the benefits and pitfalls of performance assessment (Cizek, 1991; Wiggins,

1989). Education experts discussed issues such as validity, reliability, interpretations, equating of various performance assessments, scoring and generalizability of the assessment (Kane, Crooks, & Cohen, 1999). Performance assessment can be defined as assessment made on a task which is a main emphasis of the interpretation or has high-fidelity to that task (Kane et al., 1999). Within this definition Kane et al has described two assumptions: (a) the score will coincide with a level of skill in that particular domain and (b) that the observations will consist of tasks being measured in the domain of interest (1999). The value of performance assessment is its relevance to the domain being measured (Messick, 1994). While it may be easy for some to say that a performance assessment is representative of the domain being examined, it is important to ensure that the generalizations made from an assessment are appropriate (Messick, 1995). In this study, the O-SCORE is used to make assessments of residents in a real-life surgery. In this case, the construct is surgical competency. Although most surgeons are assumed to understand what a competent trainee does, there are currently no standardized criteria for this in Plastic Surgery. Standardization of assessment is one way to improve the validity evidence for the interpretation of the assessment (Kane et al., 1999). This standardization allows for better consistency of the assessment, leading to improved reliability which is one source of validity evidence (Downing, 2003). The American Educational Research Association has defined the five sources of validity evidence as (a) content, (b) response process, (c) internal structure - under which reliability is a component, (d) relationship to other variables, and (e) consequences (*Standards for educational and psychological testing*1999). Internal structure pertains to the psychometric properties of the assessment items (Downing, 2003). It is the goal of this section to highlight the main elements of performance assessment and its reliability characteristics.

The nomenclature for performance assessment has been overtaken in the medical literature as workplace-based assessment (WBA). It has received significant attention in the past 10 years. The evidence for the use of WBA however yields mixed results as to its validity and reliability (Shalhoub, Vesey, & Fitzgerald, 2014). Before examining the body of literature on WBA an understanding of the assessment model is necessary. Miller's pyramid is often cited for assessment of technical skills (Crossley, Humphris, & Jolly, 2002; Miller, 1990). In this framework, learners progress through four stages of development. These four stages are (1) knows (2) knows how (3) shows how, and (4) does. As most surgical residency programs are five years in length, trainees require adequate exposure to surgery in order to advance through these learning stages. By understanding this framework, assessment can be developed to be aligned with the construct which is being tested. In 2011, Crossley, Johnson, Booth and Wade explored reliability when assessment scales better aligned with development of competence. This study compared assessments using three common assessment tools which consists of traditional scales of satisfactory performance to new construct aligned scales. The authors showed that raters were more consistent with their ratings and were better able to discriminate between trainees at different levels. The O-SCORE uses a form of construct aligned scale known as an entrustability scale. The authors define these scales as "behaviorally anchored ordinal scales based on progression to competence" (Rekman, Gofton, Dudek, Gofton, & Hamstra, 2016). The benefit of this type of scale is that it makes logical sense to evaluators. The assessments of clinical readiness for independent practice are based on the trust a clinical preceptor has that the trainee can manage the task at hand (Crossley & Jolly, 2012; Kennedy, Regehr, Baker, & Lingard, 2008). It is suggested that since raters based decisions in the operating room on trust, there is less of a need to translate their assessment into a different scale, potentially decreasing

measurement error (Rekman et al., 2016). Some challenges with assessment in CBME can potentially be alleviated by using construct aligned scales such as the O-SCORE entrustability scale. Another construct-aligned scale is the Zwisch scale (George et al., 2014). This scale transforms Miller's pyramid of skill development to a scale of supervision. This supervision scale is based indirectly on trust developed between surgeon and trainee (DaRosa et al., 2013). The four levels of performance on the Zwisch scale are (1) show and tell, (2) active help, (3) passive help, and (4) supervision only. Although similar to an entrustability scale, the descriptors for the levels of this scale focus on what the trainee did during the surgery versus the amount of intervention required by the rater. Although it has shown to highly correlate with the O-SCORE, the rating is a single score and does not provide a breakdown of the assessment. An itemized list of evaluative criteria can be useful in the delivery of feedback to trainees. Nonetheless, entrustability scaled items are being implemented and offer a unique option for surgical programs.

In addition to offering construct aligned scales, entrustability scales are also criterion referenced. As training programs will be more attuned to competency, criterion based assessment is a necessity. Current Plastic Surgery assessments are norm referenced. This assessment type ranks individuals undergoing the test and does not allow for interpretation of the score beyond that of those being assessed (Ricketts, 2009). It can be used for resident selection and hiring processes but its ability to properly assess competence is limited. In contrast to norm-referenced tests, criterion-referenced tests compare a trainee's performance on a task directly to the criterion for the assessment (Turnbull, 1989). In the case of CBME, competency may be the criterion for the assessment. With proper tools in place to highlight areas of performance where residents can improve, this form of testing will allow feedback for trainees. Interestingly, norm referenced

assessment is still being used in Plastic Surgery programs today despite the known pitfalls with this type of interpretation. Knowing the limitations of both norm-referenced and criterion-referenced assessment will aid program administrators to make more informed decisions regarding their selection of assessment methods.

A good assessment is more than one which yields valid and reliable scores. The Ottawa guidelines for effective assessment in medicine were developed by an expert panel who stated that there are seven elements of assessment which are important. These are (a) validity, (b) reproducibility, (c) equivalence, (d) feasibility, (e) educational effect, (f) catalytic effect, and (g) acceptability (Norcini et al., 2011). The authors of these guidelines acknowledge that criteria for good assessment depends on the nature of the assessment. Regardless, program administrators should be able to discern between assessments of varying quality.

With regards to workplace based assessment, tools using an entrustability scale offer unique advantage over traditional assessment scales. They are aligned with the entrustment, an important concept in CBME. Ten Cate first described entrustable professional activities (EPAs) that can be used in CBME to create an assessment matrix combining skills and core competencies (ten Cate, 2005). EPAs are defined as "professional activities that together constitute the mass of critical elements that operationally define a profession" (ten Cate & Scheele, 2007b). There are eight conditions of entrustable professional activities. One of these conditions is that it "is part of essential professional work in a given context" (ten Cate & Scheele, 2007b). Program administrators must now find assessment methods which are capable of assessing EPAs. This highlights one of the largest gaps in the literature, specific assessment methodology for CBME. There is still research required to determine the best way to make judgements based on assessments conducted during the course of a training program. Authors

rarely tackle this complicated subject in the plethora of papers describing CBME. It is important to develop a framework and then determine how it shall be implemented, which appears to be what has happened with CBME. Nonetheless, some progress has been made in exploring assessment in this changing environment. The next section will review some of the pertinent assessment tools available to Plastic Surgery training programs.

Operative Assessment tools for Plastic Surgery

To determine whether valid tools for assessment in Plastic Surgery exists, a thorough review of the literature was needed. Eight tools were identified. Assessment tools were excluded if they consisted of non-Plastic Surgery specific checklists, items pertaining to laparoscopic surgery, no items on technical performance, or focused on microsurgery. There is a thorough two-part systematic review of microsurgery assessment tools available in the *Journal of Surgical Education* (Dumestre, Yeung, & Temple-Oberle, 2014). This section will discuss the available assessment tools, their methodologies and the applicability to competency-based training.

Four of the articles focus on general skills for surgical procedures. These tools include the O-SCORE, OSATS, 360 degree Plastic Surgery evaluations, and the global rating index of technical skills. Of these, only the O-SCORE used generalizability theory in their statistical analysis.

Ottawa surgical competency operating room evaluation.

The O-SCORE was developed in 2012 for use in orthopedic surgery training programs but the goal of the authors was to “develop a succinct surgical assessment tool that could be used to evaluate competence on any surgical procedure” (Gofton et al., 2012). The development of this tool was based on criterion-referenced evaluation rather than on norm-referenced evaluation.

This was done to avoid a central scoring tendency where raters avoid the extremes of a rating scale (Williams, Klamen, & McGaghie, 2003). In order to examine reliability of the scores, the authors used a generalizability analysis. They included fully crossed as well as some nested facets. These facets included surgical specialty which was crossed with items, as well as observation nested within trainee which was nested within surgical specialty. In this study, where 34 surgeons evaluated a total of 163 surgeries in 37 residents, the O-SCORE had a generalizability coefficient of 0.8. The authors reported the O-SCORE assessments had evidence of construct validity. It was found the evaluations were able to distinguish between trainees at different levels training. A follow-up study was conducted to answer two questions (MacEwan, Dudek, Wood, & Gofton, 2016). The first was how do raters affect the reliability of the score? The second was how scores on the O-SCORE compare to scores on OSTATS global rating scale and checklists? They showed that an inter-rater agreement, using Cohen's kappa, was 0.89. It was also demonstrated that the O-SCORE evaluation had a very high correlation, $r(0.96)$, $p < 0.001$ with the OSATS global rating evaluation (MacEwan et al., 2016).

There is evidence that the O-SCORE can serve as a reliable assessment tool in orthopedic surgery. The main construct being assessed using the O-SCORE is thought to be competency and as such may serve as a useful tool in CBME (See Appendix).

Global rating index for technical skills.

The second general skills assessment is the Global Rating Index for Technical Skills (GRITS). Published in 2007, this global rating scale explored its usefulness in general surgery training at the University of British Colombia (Doyle, Webber, & Sidhu, 2007). The statistical analysis was conducted using a one-way ANOVA comparing mean scores at different levels of training, Pearson's correlation, and Cronbach's alpha. It was found that the mean scores were

significantly different between trainees at different years of training. Also, the mean GRITS scores and the training year were highly correlated. Cronbach's alpha, where three or four procedures were assessed, were 0.91 and 0.94 respectively. This tool was however found to have a tendency for raters to avoid using the lower boundaries of the scale. This is one of the confounding variables of the GRITS scale which was not present in the O-SCORE tool. With a focus on CBME, this tool offers an option for providing feedback however the ultimate construct of competency is not explicitly measured.

Objective structured assessment of technical skill.

One of the most widely cited surgical assessment forms is the Objective Structured Assessment of Technical Skill (OSATS). This assessment was developed at the University of Toronto and has shown validity evidence in general surgery training (Martin et al., 1997). Other assessment tools, such as the GRITS described above, based its development on the structure of the OSATS (2007). This is an assessment tool which was evaluated using live animal models as well as bench models for six general surgery procedures. The scoring system is divided into three components. The first a task-specific checklist, second is a seven-item global rating scale and the third aspect is a pass/fail judgement. In order to examine inter-rater reliability, the authors used intraclass correlation coefficient. Evidence of construct validity was assumed to be present when significant score differences are seen between trainees of varying training years. In this initial study of validity and reliability of the OSATS the authors found that year of training did account for a significant proportion of the variance between scores. Cronbach's alpha scores however were low; ranging from 0.33 to 0.74. This measure of internal consistency is lower than what is seen using other tools. The authors state that a calculation using the Spearman Brown Prophecy

Formula showed that including two additional skills stations would increase the reliability of the global rating scale to 0.8.

The use of OSATS and its validity has been analyzed in other centers and specialties. The face validity of the OSATS was evaluated by obstetrician and gynecologists in the United Kingdom (Bodle, Kaufmann, Bisson, Nathanson, & Binney, 2008). They found that 76-80% of assessors and trainees thought the OSATS was able to assess surgical skills. As well, 76 % of the respondents felt that OSATS should be incorporated into their training program. To explore geographical differences in assessment performance, the OSATS was implemented and evaluated in Japan (Niitsu et al., 2013). The authors of this Japanese study evaluated trainees using only the global rating scale portion of the OSATS after dividing the surgical cases into those where the trainee was the first assistant or whether they acted as lead surgeon. Although using a different statistical approach than the original authors of the OSATS, this study found the scores increased significantly with each year of training. The OSATS was also evaluated in dermatology (Alam, Nodzenski, Yoo, Poon, & Bolotin, 2014). They found that there was no association between length or quality of dermatologic surgery training and the scores on the OSATS. Again, in this study all trainees did relatively well and the entirely of the scale was not seen used by the assessors. This is perhaps the case where all trainees perform exceptionally well. These studies do not discuss whether the assessors are basing their scores relative to trainees at a similar level or based on a set benchmark independent of training year.

Clarification of the scale judgements made using the OSATS may better the argument for use in CBME. If the scale is based on overall competency rather than based on rating comparative to level of training, then it may be more appropriate in a CBME. Overall, there is validity evidence for using the OSATS but better construct alignment is necessary.

University of Kentucky division of Plastic Surgery resident evaluation form.

The fourth general evaluation tool for Plastic Surgery is the University of Kentucky division of Plastic Surgery Resident Evaluation Form (Pollock, Donnelly, Plymale, Stewart, & Vasconez, 2008). This is a four-part evaluation which examines elements in addition to surgical ability. This evaluation can be considered as an overall assessment for summative evaluation in addition to surgical skills assessments. Interestingly, this evaluation tool was able to elicit a difference in evaluation ratings of trainees on operative/technical skill when completed by surgeons but not when nurses completed the evaluation. This highlights an important point. If non-surgeon health care professionals are to be included in assessment, then the tools that are used must be able to elicit a difference between groups, otherwise critical measurement errors may occur.

Task specific operative checklists.

Although there is a trend to use global-skills evaluations, task specific assessments may still be required. It has been discussed that further development of objective tools in the assessment of Plastic Surgery trainees is needed (Knox et al., 2014). Recently, a methodology for efficient development of task-specific checklists was piloted in Plastic Surgery (Courteau, Knox, Vassiliou, Warren, & Gilardino, 2015). The result was two assessment tools which can be used to evaluate face lifts and breast augmentation. They introduced the notion of the Delphi technique to reach a consensus of 90% amongst surgical experts. After two rounds of surveys to gather opinions on included items, Cronbach's alpha was found to be 0.87 and 0.85 for the face lift and breast augmentation evaluation forms respectively. These evaluation tools have not yet been studied to show construct validity.

Combined knowledge exam and directly observed performance evaluation.

Upper limb surgery is performed by Plastic Surgeons as well as orthopedic surgeons. VanHeest et al (VanHeest et al., 2012) examined the use of task specific checklists and global rating scales, as in OSATS, for evaluation of common upper limb surgery. They have published two articles which attempt to show valid and reliable use of a modified OSATS evaluation. For use in carpal tunnel surgery they evaluated residents using a three step process (Van Heest et al., 2009). The first was a knowledge examination which was designed by a hand surgery expert and underwent three iterations prior to use. Residents performed one carpal tunnel release and two examiners evaluated their performance using a task specific checklist, a global rating scale and a pass or fail judgement. Cronbach's alpha was 0.75 for the task checklist and 0.82 for the global rating scale. It is interesting to note that of the trainees who scored less than 70 (out of a possible 100) on the knowledge examination, no individual passed the practical portion. Scoring greater than 70 however did not guarantee that the trainee passed the practical examination. This further emphasizes the need for objective surgical skills assessments as the current model for licensure involves only written and oral examination. Without the use of technical performance assessments trainees may not be surgically competent, despite being successful on the summative knowledge based licensure examinations. In Canada it is the responsibility of the program director to ensure adequate assessment of the candidate's technical ability.

A second upper limb evaluation by VanHeest et al (2012) used detailed checklist and a global rating scale to evaluate orthopedic surgery residents in three surgical procedures: trigger finger release, carpal tunnel release and distal radius fracture fixation. This was evaluated using cadaver stations as high fidelity simulation. The detailed checklist was created by two board certified surgeons with expertise in hand surgery. Their analysis showed that there as no

evidence of construct validity for surgical procedures using this assessment tools. They also failed to show acceptable interstation reliability, Cronbach's alpha was 0.43 for the detailed checklist and 0.56 for the global rating scale.

This form of assessment is valuable for teaching purposes but the authors do not show how it compares to other assessment methods.

Comprehensive observations of resident evolution.

Recently, a group in Baltimore has published results of their comprehensive observations of resident evolution (CORE) assessment tool (Cooney et al., 2016a). This assessment is a web-based tool, which asks raters to evaluate trainees on their operative performance. It is based on the operative entrustability assessment (OEA) (Cooney, Redett, Dorafshar, Zarrabi, & Lifchez, 2014). This OEA evaluates residents in a manner similar to the O-SCORE assessment scale. At one end of the OEA scale is that the resident would not be able to perform the surgery and at the other end the resident is deemed capable of performing the surgery independently. The authors have provided validity evidence through a regression showing increases in scores based on year of training. There is no published reliability evidence for this specific assessment. This tool is aligned with competency-based training and offers significant potential as a useful assessment method.

Overall, there exists a discrepancy in the validity and reliability of assessment tools available for use in Plastic Surgery. In 2011, a US-based Plastic Surgery Milestone project working group was assembled to develop a structure which would guide the speciality toward evaluation based on objective measures of competency (McGrath, 2014). There are evaluation forms available on the ACGME website however there is no published evidence indicating its

validity or reliability. Additionally, in Canada formative and summative evaluations are conducted using In-Training Evaluations of Residents (ITERs). These are based on the Royal College of Physicians and Surgeons of Canada CanMeds framework. However, the evaluation forms in use have not been evaluated in surgical residents for construct validity or reliability. The development of an ITER for emergency medicine residents was performed using exploratory factor analysis which showed that a 24 item evaluation could be used to measure the desired abilities individually with Cronbach's alpha of greater than 0.9 for each subscale used (Kassam, Donnon, & Rigby, 2014). It is imperative that an effort is made to develop and validate surgical skills assessment tools which can be used to guide promotion criteria, ensure public safety and enhance the training of surgical residents.

Reliability of assessment

Reliability represents the ratio of true score variance to the amount of true score and error score variance. The operational definition of reliability varies. It may be defined as the reproducibility of a measurement or the ratio of true score variance to true score and error variance (Haertel, 2006). Ultimately, the more error variance, the less reliable that measure becomes. Reliability is important to the development of validity evidence (Messick, 1980). This is due to the fact that it is difficult to claim an assessment is measuring the intended construct when it is not reproducible. This can be illustrated by a simple example. If a rater is asked to classify trainees into two categories but is only correct 50% of the time, the decision being made is not reliable. It is unlikely that any decision from that rater is valid as the interpretation of the assessment is not reproducible beyond a chance agreement. An important concept when examining reliability in assessment is the validity reliability paradox (Kane, 1982). This represents the balance between the breath of a domain being measured and the effect on validity

of the assessment. As a domain is narrowed, the reliability of the assessment measure may increase however, the validity may decrease (Brennan, 2000). This is due to the decrease in error variance offered by the domain and the decrease in generalizability of the assessment. Throughout this section aspects of reliability applicable to this study, (a) classical test theory, (b) internal consistency and (c) rater reliability will be described.

Classical test theory.

First described by Spearman in 1904, Classical Test Theory (CTT) still serves as the basis for many measurement procedures used today (Brennan, 1992; Spearman, 1904). It is the classic notion that an individual's true score (τ) is equal to the observed score (X) plus an element of measurement error (ε).

$$\tau = X + \varepsilon \quad (1)$$

Classical Test Theory is accompanied by eight assumptions. The in-depth review of these assumptions are not necessary for an understanding of Generalizability Theory and will not be discussed. To use CTT in evaluation of assessment tools does allow for an analysis of error associated with the tool but it does not partition error variance. To begin a discussion of reliability, it is important to note that under CTT the true score cannot be correlated with the error term (Traub & Rowley, 1991). Reliability is mathematically represented by equation (2) which states it is equal to a ratio of the true score variance to the observed score variance.

$$\rho_{xx'} = \frac{\sigma_\tau^2}{\sigma^2_{X=\sigma_\tau^2+\sigma_E^2}} \quad (2)$$

In this equation, $\rho_{xx'}$ represents the reliability coefficient, σ_τ^2 is the true score variance and $\sigma^2_{X=\sigma_\tau^2+\sigma_E^2}$ is the observed score variance, a combination of true score and error score. The possible values for a reliability coefficient range from zero to one. A reliability value close to one indicates the majority of the variance is in the true score and not attributed to measurement error. As reliability decreases, there is assumed to be greater measurement error. This is not always the case and will be described later in this section. The reliability of any test is said to be dependent on three factors: (a) the test, (b) the conditions, and (c) the examinees (Traub & Rowley, 1991). The number of items on a test is known to affect the reliability of the scores (Haertel, 2006). In general, more items will result in higher reliability. This is because the error resulting from item variance will decrease as the number of items increases. Spearman-Brown prophecy formula can be used to determine the reliability if the number of items were to change (Spearman, 1910). This is very useful for optimizing assessment methods. The equivalence to this in generalizability theory, which is presented later, is a decision study. Thus far, the grounding principle of CTT has been shown. The true score represented in the CTT formula is not actually able to be measured. As such, to calculate the reliability coefficient repeated measures are used.

One method of using Classical Test Theory to calculate reliability is test-retest. In this form, a test is administered one two separate occasions. It measures the degree of stability in the measurement. In this form of reliability testing, the reliability coefficient is a correlation between the two test scores. A reliable test would have a high correlation between the two administrations of the test. There are logistical implications for this type of study design. One such problem is the effect of recall and testing conditions on the subsequent score once a candidate has been exposed to the test. In a high stakes examination, such as a medical licensing exam, this type of

methodology would not be feasible. Test retest is not a viable option in performance based testing in residency as the abilities of the trainees are likely to have changed in the time between assessments. An example of this could be in a simulated cadaver operation which is repeated. If test-retest was used for the same surgery, one months apart, it is likely the resident's skill would have changed in that time. The reliability is no longer assessing the score appropriately as the trainee's ability has changed. Another method, such as generalizability theory, may be more appropriate.

To overcome these issues, alternate forms can be used. This method requires the development of two tests which are deemed to be assessing the same construct at the same level of difficulty. Creating an additional analogous test is time consuming and potentially expensive. As with test-retest there are aspects of the conditions of measurement, such as fatigue of examinees, which contribute to the error between two administrations. This form of reliability testing can be used in residency training for knowledge tests as well as in performance-based assessment.

In most cases, the ability of residency training programs to perform assessment is limited. The next section will discuss two measures of reliability used in this study.

Types of reliability.

It is important for those using reliability data to understand exactly what the value being used actually describes. Reliability coefficients can take various forms and represent different concepts depending on how the coefficient's calculation was performed (Traub & Rowley, 1991). For example, measuring internal consistency provides different information than does

measuring inter rater reliability. This section will highlight the differences between internal consistency and inter-rater reliability and methods for their calculation.

Internal consistency.

Internal consistency is the degree to which scores across items are stable (McGoey, Cowan, Rumrill, & LaVogue, 2010). Moreover, internal consistency can be viewed as how scaled items interact to measure a similar construct. Lee Cronbach described the determination of a reliability coefficient for single measures of items scored with more than two values (Cronbach, 1951). Equation below illustrates how alpha is calculated (Cronbach & Shavelson, 2004)

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum s_i^2}{s_t^2}\right) \quad (3)$$

In the above equation (3), k represents the number of items, s_i^2 is the standard deviation of the item scores and s_t^2 is the total score standard deviation. The value is the ratio of test item covariance to the total test variance. It has also been called the internal-consistency coefficient (Webb, Sharvelson & Hartel, 2006).

There are three important facts regarding alpha and its results. The first is that in theory it is described as the lower bound estimate of reliability and as such, is a conservative estimate of reliability (Carmine & Zeller, 1979; Novick & Lewis, 1967). This however assumes that the items are parallel, meaning they "measure the same ability at the same level of difficulty" (Traub & Rowley, 1991). The second is that alpha is dependent on the number of items. From equation (3), as the number of items increases so does the $\left(\frac{k}{k-1}\right)$ value. The third is that alpha can be used as evidence for unidimensionality (Tavakol & Dennick, 2011). This concept important as alpha

inflation can occur when large number of multidimensional items are included in the calculation of alpha. If certain items are known measure particular traits, than the alpha for each sub section should be reported.

Cronbach's alpha is widely used in medical education literature and readers should be aware what information this is providing. A single administration of a test allows for use of Cronbach's alpha as a measure of internal consistency.

Inter-rater reliability,

Raters are often a source of measurement error in assessment of medical trainees (Swanson & Others, 1995). Factors affecting rater consistency include training, familiarity with the tool, standardization, and rater bias toward the trainee (Hallgren, 2012). Regardless of these factors, it is important to understand how raters affect the measurement and the associated error. This section will review intraclass correlation (ICC) to calculate inter-rater reliability.

There are two trends in the measurement of inter-rater reliability. These are calculation of Cohen's kappa and intra-class correlation dependent on whether or not the variable is nominal. Cohen's kappa is a measure that determines the agreement between raters with a correction for chance agreement (Cohen, 1960). It is used in cases were assessment is on a nominal scale. As most assessment scales currently used in medical education are ordinal, the use of intra-class correlation is often more practical.

Intra-class correlations are based on ANOVA models (Gisev, Bell, & Chen, 2013). In 1979, six forms of ICC were described (Shrout & Fleiss, 1979). These are defined based on the study design. An ICC can be defined as a ratio of the variance of interest and the sum of the variance of interest and error. The result from calculating an ICC is the determination of the

variability provided by the subject and not from rater error. An ICC can have a value from zero to one. As the ICC value approaches one, the error in measurement due to raters is minimal. Overall this offers a useful measure for determining inter-rater reliability.

Generalizability theory

As discussed in prior sections, the measurement associated with assessment is not without error. Classical test theory states that a true score is the sum of the observed score and an error term (Brennan, 1992; Spearman, 1904). Generalizability theory can be viewed as a method of separating the error term into multiple error components (Cooper & American Psychological Association, 2012). It combines elements of Classical test theory and ANOVA procedures. This section will explore generalizability theory as it pertains to evaluation of assessment tools.

One of the described benefits of G theory is that it is said to "liberalize CTT by permitting a decomposition of the undifferentiated error term into multiple parts" (Brennan, 2000, p. 339). The ultimate structure of using G theory is to design a study with the desired sources of error included, determine the variance components through ANOVA measures and then determine generalizability coefficients. Once variance components are determined a decision study can be conducted to look at changes in reliability when test administration is changed.

Nomenclature in G theory.

Using generalizability theory requires familiarity with some unique nomenclature. The first is that a facet is defined as a source of error (Streiner, Norman, & Cairney, 2014). Facets chosen represent potential sources of error when determining the generalizability of the measurement. The investigator must choose facets for which to examine variance components. In

a study of workplace based assessment in surgery, facets could be rater, items, occasion and surgery type. When using generalizability theory it is up to the investigator to design a universe of admissible observations (Shavelson & Webb, 1991). "Admissible" is what the investigator views as acceptable to the study and the population. The term universe is used to discuss conditions of measurements. One of the most important uses of the term universe is the universe of generalization. This represents the universe to which the investigator wants to generalize the results of a particular study. It is important to note that the broader the universe of generalizability the more likely it will be to commit error when generalizing from a sample to the universe. Narrowing a universe of generalizability can result in increased reliability due to the decrease in error.

Generalizability studies.

A generalizability study examines the variance components for the various facets. These variance components can be used to conduct a decision study (D study). A D study can model the changes in the variance components when the study design is modified, such as an increase in the number for items used in the assessment. This can be thought of as the analogue to the Spearman Brown Prophecy formula in CTT (Webb, Shavelson, & Haertel, 2006). The designs of these studies can vary in complexity. The facets used can be crossed or nested, which increases the mathematical complexity when using G theory. Crossed facets occur when any combination of the two facets would be admissible to the investigator (Brennan, 2001). It can also be considered crossed when all conditions of one facet are seen with all conditions of another facet (Shavelson & Webb, 1991). An example of this would be a study where all raters examine each candidate. In this case, every candidate will be seen by each rater. Nested facets occur when two or more conditions of the nested facet appear with one and only one condition of another facet.

An example of this in surgery assessment would be to have one rater evaluate the residents on one occasion and a different rater for the second. In this case, rater is nested within occasion. Finally, it is important to discuss random versus fixed facets as, mathematically, the variance components are handled differently. Random facets represent a sample of all admissible conditions for that facet. For example, if there are 100 Plastic Surgeons capable of performing evaluations but only two are chosen at random for the study, then rater (or surgeon) is considered a random facet. Fixed facets occur when all conditions for that facet are observed in the G study and generalizing beyond them is not logical or of benefit to the investigator. In other words, the entire facet is employed in the generalizability study and does not represent a random sampling. An example of this is when using a pre-designed checklist for assessment. If item is chosen as a facet and the items are unique to a specific evaluation tool, then the investigator cannot generalize beyond those items which are on the assessment form. In this case the facet item would be considered as a fixed.

As stated earlier G theory is a combination of elements from CTT and ANOVA. Essentially, generalizability theory uses a factorial ANOVA to derive the variance components for the G study. In this analysis, an estimate mean square is calculated for each facet and the interaction terms. Each facet chosen will offer a 'main effect' variance and variance of the interactions with the object of measurement and/or the other facets. Take the example of a person being evaluated on 10 items and over multiple occasions. If in this case the design is crossed, then there will be a main effect for the person taking the test as well as the occasion and the items. There will also be an effect of the interaction between these facets, as well as a residual error variance. If this error variance is high, or accounts for a significant proportion of the total variability then it can be said that potentially, then there are other facets not included in the study

design which are influencing the variability. In this example there are seven components of variance which must be analyzed when conducting the G study. A Venn diagram highlighting the variance components when a crossed, two facet design is shown. If in the above example, each person was evaluated on different occasions using the same items, than occasion would be nested within person. This is represented in Figure 1.

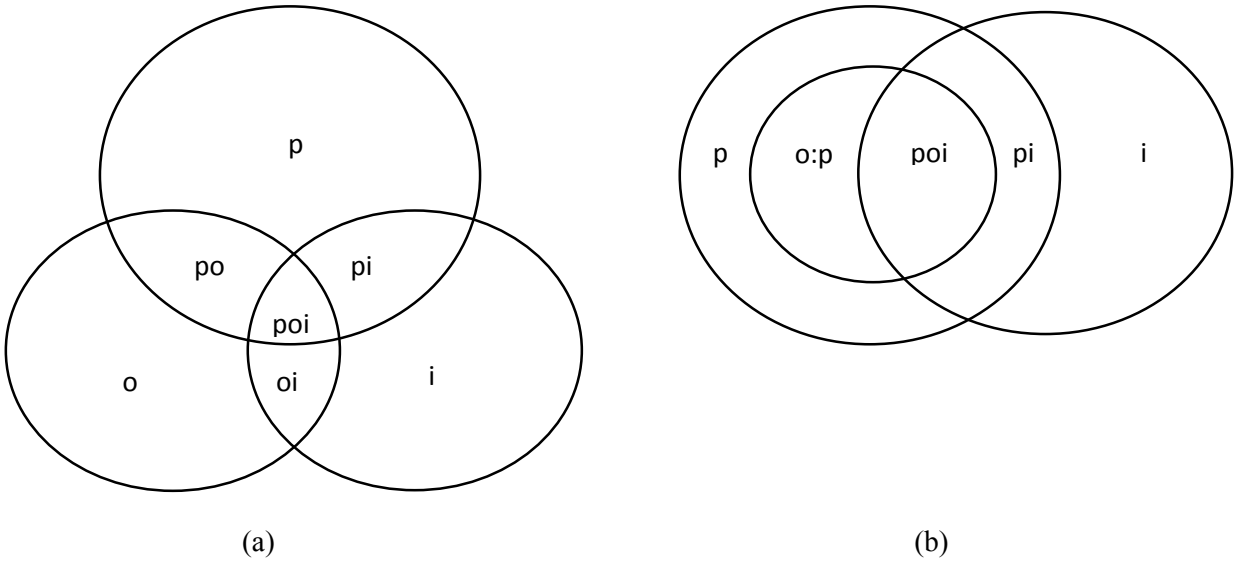

Figure 1. Venn diagrams depicting (a) the sources of variability for a fully crossed, two facet, G study (*p* x *o* x *i*) and (b) the sources of variability for a nested *o*:*p* crossed with item design (*o:p* x *i*)

Once the variance components are found in the G study, these are used to make inferences about the test. These variance components however do have limitations, depending on the initial study design. The most widely useable and informative variance components are from a crossed design and the main effect for each facet is known. When facets are nested, the

investigator is not able to fully separate the variance components of the facet nested within another (Shavelson & Webb, 1991). For a G study in this case, rater would be nested within trainee (*r:p*) The main effect of the rater would be confounded within the interaction between rater and trainee as well as found in the error term. Using these variance components a D study will typically examine the relative and/or absolute error variance, the generalizability coefficient and the index of dependability.

Decision studies.

The relative error variance is a combination of a weighted sum of all the variance components which have interactions with the object of measurement. The variance components which do not interact with the object of measurement do not contribute to the error associated with the relative standing of persons. The absolute error variance however, includes a weighted sum of all the variance components except the universe score variance. Because the absolute error includes more variance components in its determination, it is typical that absolute error variance is larger than relative error variance. The generalizability coefficient in G theory includes the relative error variance in its calculation. This is defined as the "ratio of universe score variance to itself plus the relative error variance" (Brennan, 2000). It is thought to be the CTT analogue to a reliability coefficient (Gao & Harris, 2012). It is defined by equation four (Brennan, 2000).

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\delta)} \quad (4)$$

In this equation Ep^2 is the generalizability coefficient, σ^2 (τ) is the universe score variance, and $\sigma^2(\delta)$ is the relative error variance. The dependability coefficient is based on the absolute error

variance, $\sigma^2(\Delta)$, and is represented the letter phi (Φ). This reliability coefficient is represented by:

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} \quad (5)$$

Overall equations (4) and (5) represent the reliability coefficients which are the ultimate outcome from a D study. The decision to use relative versus absolute error is dependent on the research question. For rank ordering individuals, a norm referenced method of assessment, the relative error should be used. For decisions on overall ability irrespective of the sample being evaluated absolute error should be used. In the case of competency-based assessment, rank ordering individuals is not adequate in the determination of true competency. All possible sources of error should be included in the determination of reliability of such assessment scores.

Generalizability theory serves as a useful methodology to assess reliability. Studies must be appropriately designed to determine the variance components of interest. In competency-based training, assessment is a major focus. This is due to the underlying reason for a shift toward CMBE, accountability to the public. As program administrators will be responsible for the implementation of assessment, they must also be aware of the properties of validity and reliability of the scores which they are using to make high stakes decisions of medical trainees.

Methodology

Ethical Considerations

The University of Alberta Research Ethics Board approved all parts of this study. Consent for trainees was collected by the investigator (CRB) prior to data collection.

Sample

Video analysis.

Plastic Surgeons in attendance at a local divisional meeting, where all Plastic Surgeons are invited and encouraged to attend, were selected to participate in this part of the study. In total, seven Plastic Surgeons participated. These participants were all fellows of the Royal College of Physicians and Surgeons of Canada in Plastic Surgery. All participants have experience training and evaluating surgical residents.

Generalizability Study.

Plastic Surgery residents at the University of Alberta were invited to participate in the study. There were fourteen residents enrolled in the training program at the time of the study. Of these, twelve residents were eligible. Two residents were excluded from the study because they were not performing regular clinical duties and would be unable to participate. The Plastic

Surgery residents were all Canadian medical school-trained physicians. They ranged in post graduate training year (PGY) from one to five. The raters for the study were Plastic Surgeons in Edmonton, Alberta affiliated with the University of Alberta. These surgeons are considered experts in the domain of Plastic Surgery and are all Fellows of the Royal College Surgeons of Canada. In total 13 Plastic Surgeon raters participated in the workplace-based assessment using the O-SCORE.

Procedure

This study comprised of three parts. The first was a reliability study to examine internal consistency and inter-rater reliability. This part utilized two data sets. The first was a blinded video analysis to score technical ability using the O-SCORE. The second data set was the workplace-based assessment using the O-SCORE tool. The next part of the study explored validity evidence. Scores of the O-SCORE assessment based on year of training were determined. The differentiation of trainees based on year of training was used as evidence of validity. The third part of the study was a generalizability study based on eight months of evaluations of Plastic Surgery resident performance in the operating room. The methodology for each of these three parts are described in the following sections.

Part 1- Reliability Testing.

This part of the study examined the internal consistency of the items and interrater reliability using video and workplace-based assessment.

Video development.

A video was created to obtain data on internal consistency and interrater reliability. After informed patient consent, an open reduction internal fixation of a metacarpal fracture was video

recorded at the Royal Alexandra Hospital. This procedure was selected as (a) the resident often acts as primary surgeon with assistance from the attending staff person, (b) it is a procedure which is included in the generalizability study, and (c) it is a common Plastic Surgery procedure. It was recorded by the study author, CB, over the shoulder of the operating surgeon. A Sony Handicam, HDR-PJ380 was used to record the surgery. A third year Plastic Surgery resident at the University of Alberta performed the surgery with guidance and intermittent operating by the attending surgeon. The total recorded video time was 54 minutes. The staff surgeon intervened during the reduction of the fracture, approximately 10 minutes of operative time, otherwise the procedure was performed by the resident. The film was edited to yield a three minute surgical video of all the key steps of the procedure. Editing was performed using Adobe Premier Elements 10 [computer software]. The key surgical steps included skin marking, skin incision, fracture exposure and reduction, internal fixation, closure and splint placement. These were all shown in the video. Despite being edited, it was important to capture all aspects of the surgical procedure in order for evaluators to gain perspective on overall ability. Where any performance step was repeated multiple times, such as drilling holes and placing screws, only the first was shown in the video. At the stage where staff surgeon performed the reduction a slide was inserted informing the examiners that the surgeon performed that component of the surgery. This was done to increase the likelihood that surgeons watching the video would notice the intervention required by the supervising surgeon. The video was made without audio. The video was intentionally made to be as short as possible in order to increase the participation of surgeons in our study.

Video rating.

Seven Plastic Surgeons present at the Plastic Surgery divisional meeting watched a three minute surgical video as described above. The raters were explained the O-SCORE with a two minute presentation highlighting the entrustability scale and the O-SCORE items. In the case of this analysis, only three items were evaluated. These were (a) technical ability, (b) efficiency and flow, and (c) visuospatial ability. These items were chosen as they represent constructs which could be analyzed without interacting with the trainee. These items are thought to represent the construct of technical ability. Following participant education, the video was played on a large projector screen. It was played once from beginning to end. Immediately following the video, they independently completed the assessment on a paper version of the three item O-SCORE. The scale and the item descriptors were included for reference.

Assessment tool education.

The O-SCORE tool was used as an assessment method for surgical performance. Prior to data collection, users of the tool were provided detail on the tool's format and scale. This education is thought to improve the understanding of the rating scale, the items and the inter-rater reliability.

The Plastic Surgery residents were given a 10 minute information session on the O-SCORE tool and its assessment scale. This occurred in June 2015 prior to the initiation of the data collection. During this session, residents were presented with an overview of the O-SCORE entrustability scale. The presentation included the timeline for the study and highlighted the scale and items. Participants were given the opportunity to ask questions. Through discussion residents were supportive of the study and understood the entrustability scale. Residents were explained that a low score using the O-SCORE does not equate to the scales currently used at the University of Alberta. Based on previous work by Gofton et al. (2010), it was shown that the

residents in their study did not become discouraged by low scores if they were junior trainees or inexperienced with the technique.

All surgeons who participated as raters in this study were provided with face-to-face individual training regarding the study and the O-SCORE. The study author (CB) met with each surgeon and provided an overview of the study and the assessment tool. This was done in June 2015, prior to initiation of the data collection. A paper copy of the tool was provided to them. In particular, surgeons were educated about the O-SCORE scale as it differs from other assessment tools used at the University of Alberta. In addition to this face-to-face interaction, an email summarizing the study and a copy of the O-SCORE tool was sent to all Plastic Surgeon participants.

Data analysis.

Item-total correlation, Cronbach alpha, and descriptive statistics were completed using IBM SPSS (Version 23.0). Descriptive statistics included item mean scores, standard deviation and standard error of the mean. Assumptions for statistical tests were verified and are reported in the results section.

Part 2- Validity Evidence.

The differentiation of resident level of training based on O-SCORE assessment was used to provide validity evidence in this study. The workplace-based assessment data was used for this part of the study.

Data analysis.

IBM SPSS (Version 23.0) was used to perform statistical tests in this part of the study. Initially a MANOVA analysis was performed. From this analysis it was evident that trainees could be analyzed in two groups. As such, the MANOVA was followed by multiple t-tests to

further examine differences between junior and senior residents. A chi square test was also performed to examine difference in scores on the dichotomous item assessing independence amongst junior and senior trainees. Statistical significance was set at $p < 0.05$.

Part 3- Generalizability Study.

A generalizability study (G study) was designed in order to examine the reliability of the O-SCORE assessments when used in Plastic Surgery. The workplace-based assessment data was used for this part of the study. The universe of admissions was defined by the Plastic Surgery residents, the number of occasions on which they were evaluated, and the individual O-SCORE items.

The original study design was a fully crossed, balanced, G study. The facets were surgery type, occasion and item (*s* x *o* x *i*). The plastic surgery residents were asked to be evaluated a total of six times. Of these six evaluation, two were to be for each of (a) breast reduction mammoplasty, (b) mandible open reduction internal fixation, (c) and hand fracture fixation. Two evaluations from three types of surgery was deemed adequate as it allows determination of the variance components in the G study. It was also agreed that it would be a reasonable request of busy surgery residents, voluntarily participating in the study. The surgery types were selected as they represent common plastic surgery procedures at the University of Alberta. They are also procedures where residents are typically able to perform some or all of the operating. This is an important consideration given the type of assessment scale used in the O-SCORE. As the study progressed, residents were not performing sufficient numbers of the set tasks to complete the study. The G study design was modified to allow analysis based on the data collected. The subsequent G study design, and that used throughout this thesis is a mixed, unbalanced partially-nested G study (*o:r* x *i*).

The occasion facet was nested within resident. This was necessary because residents were evaluated on different occasions. The resultant study was an item crossed with occasion within resident design (*o:r* x *i*). Occasions and residents, were deemed to be random variables as they represented a only sample of all potential observations. The item facet however, was treated as a fixed facet since these eight O-SCORE items represent all possible items in this generalizability study. The assessment is not to be generalized beyond the items of this tool. Since the design for the generalizability study includes random and fixed facets, the analysis was performed using a mixed random effects model. The G study was also unbalanced. There was missing data as some residents did not receive six evaluations.

Data analysis.

To conduct the generalizability analysis, G_String_IV (Version 6.1.1) [computer software]. Hamilton, ON; Blotch & Norman, was used. Variance components were calculated based on estimated mean squares and a dependability coefficient determined. A decision study was performed to examine the reliability coefficient when the number of occasions is changed. The range of occasions in the D study was one to ten.

Results

Part 1- Reliability testing

Item Reliability Testing

In this section, the items and overall score of the O-SCORE evaluations are examined with respect to the internal consistency and relationships between items. This was conducted using data collected from the workplace-based resident assessments as well as by using standardized video analysis.

In order to examine internal consistency, the corrected item-total correlations were determined. The results are summarized in Table 3**.** All items had high corrected item-total correlations (range 0.74 – 0.88). Knowledge of steps had the highest item-total correlation. The post procedural plan had the lowest item-total correlation. These item-total correlations are shown in Table 1.

Table 1

O-SCORE corrected item-total correlation for entrustability scaled items

Item	Item-total correlation
Preprocedural plan	0.83

Case preparation	0.85
Knowledge of steps	0.88
Technical skills	0.82
Visuospatial skills	0.85
Postprocedural plan	0.74
Efficiency and flow	0.84
Communication	0.82

The internal consistency was also examined using Cronbach alpha. The Cronbach's alpha value was 0.95. This value is very high. The conditions for measurement of this value were not standardized and as such, a simulated and blinded analysis was performed during a video analysis.

The correlation between the overall score and the dichotomous item of competence was also calculated using Pearson Product Moment Correlation. The positive correlation was $r(0.73)$, $p < 0.001$. The correlation between each of the entrustability scaled items and the competency item was calculated. The item, preprocedure plan, had the highest correlation with evaluation of independent readiness, $r(0.78)$, $p < 0.001$. This was followed by knowledge of steps, $r(0.74)$, $p < 0.001$. The item assessing communication had the lowest correlation with the independent readiness, $r(0.49)$, $p < 0.001$.

Further item analysis was performed using video analysis of a surgical procedure. Seven surgeons viewed a three minute, edited, video and rated the trainee on technical skills, visuospatial skills and efficiency and flow. The raters were blinded to the resident performing the surgery. Internal consistency of the data was again analysed using Cronbach's alpha. This was repeated under standard conditions to more accurately represent the internal consistency. Three items on the O-SCORE were assessed for internal consistency. These items were the technical

skill items which are able to be assessed using video data. The additional items were not deemed to be appropriate for video analysis. With seven raters evaluating three items, the resulting Cronbach alpha was 0.90.

An inter rater reliability analysis was performed on the video scoring data. In this case there were seven raters, one trainee and three items. Standard error of the mean and standard deviation was examined. In the case of surgeons watching a single video, the SEM was 0.36 for technical performance and visuospatial ability. For efficiency and flow, the SEM was 0.37. the standard deviation of these item scores ranged from 0.95 to 0.98.

Overall these results show that the items on the O-SCORE measure a unidimensional construct. The internal consistency measures are very high. The correlation between each item and the item assessing overall competency are significant, especially items assessing technical skills. The inter-rater reliability was low for evaluation of the three items.

Part 2- Validity Evidence

This part of the study explored differentiation of trainees based on year of training using the O-SCORE. Various methods were used to explore the data for these differences. Plastic Surgery residents of all years of training were invited to have six evaluations completed. Ten residents participated in the study. The study design was a mixed, two-facet, generalizability study. Over eight months (July 2015-January 2016) a total of 41 evaluations were completed by 13 Plastic Surgeons. Individual residents were evaluated over a range of two to six times (*M* 3.7, *SD* 1.49). In total, residents were evaluated on nine mandible open reduction internal fixation, 14 hand fracture fixations, and 18 breast reduction mammoplasties. Table 1 summarizes the distribution of evaluations based on year of training.

Table 2

O-SCORE evaluations collected per year of training

Postgraduate year	Number of evaluations
1	5
2	5
3	8
4	9
5	14

Initial evaluation was performed using MANOVA to determine if there were significant score differences between the various postgraduate year levels. This was selected in order to minimize family wise type one error that would be present if multiple individual t tests were performed. In the MANOVA analysis there were nine dependent variables. These were the eight scaled items and the overall score. There were five levels of the independent variable, the five PGY resident groups. In the multivariate analysis omnibus test, there was a significant difference between trainees based on their year of training, $\Lambda = 0.08$, $F(4,36) = 2.77$, $p < 0.001$. The differences based on postgraduate year are illustrated in Figure 1.

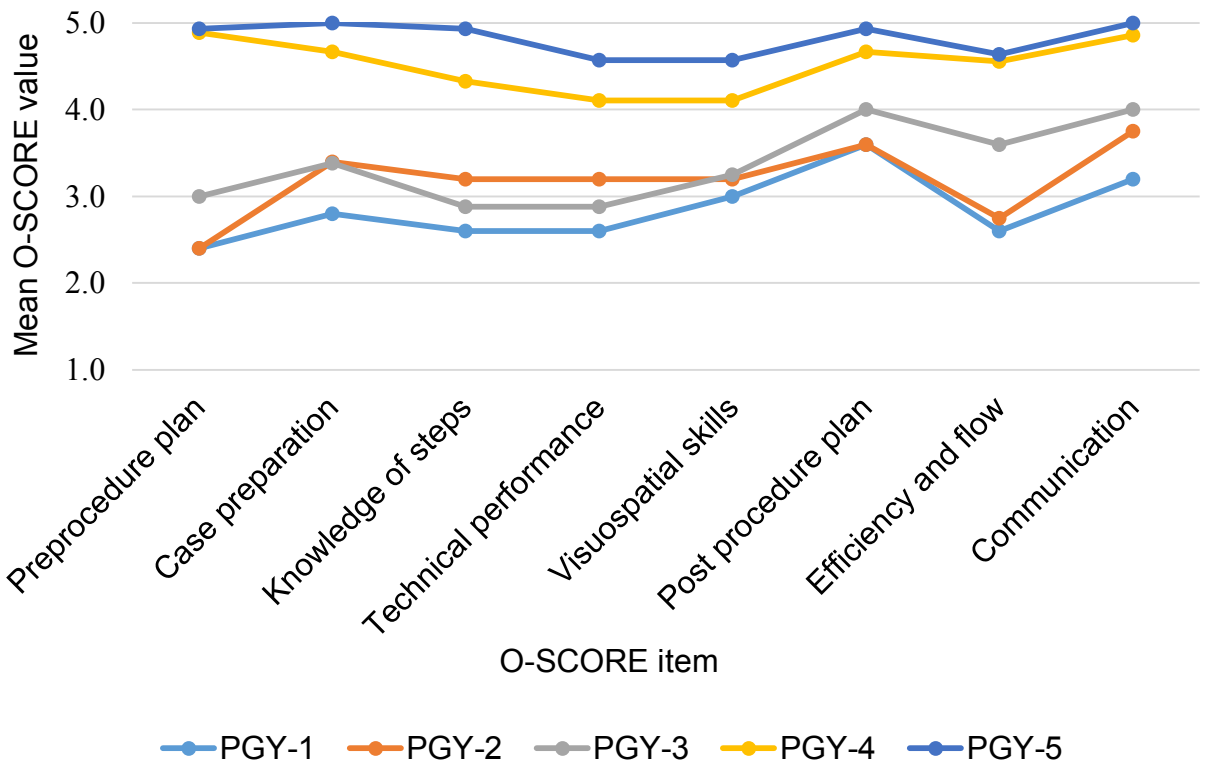

Figure 1. Mean O-SCORE item score based on postgraduate training year. Possible score on each item ranges from one to five.

This figure illustrates that there is a difference between trainees on each item based on their level of training. This difference provides validity evidence for the use of the tool to assess readiness for independent practice. It is assumed that residents would be closer to readiness for independent practice as they progress through training.

Next, Tukey post hoc test, was conducted to examine specific differences between groups. From the post hoc test it was evident that there were significant differences based on whether the resident was junior or senior. The overall scores for each year are shown in Figure 2.

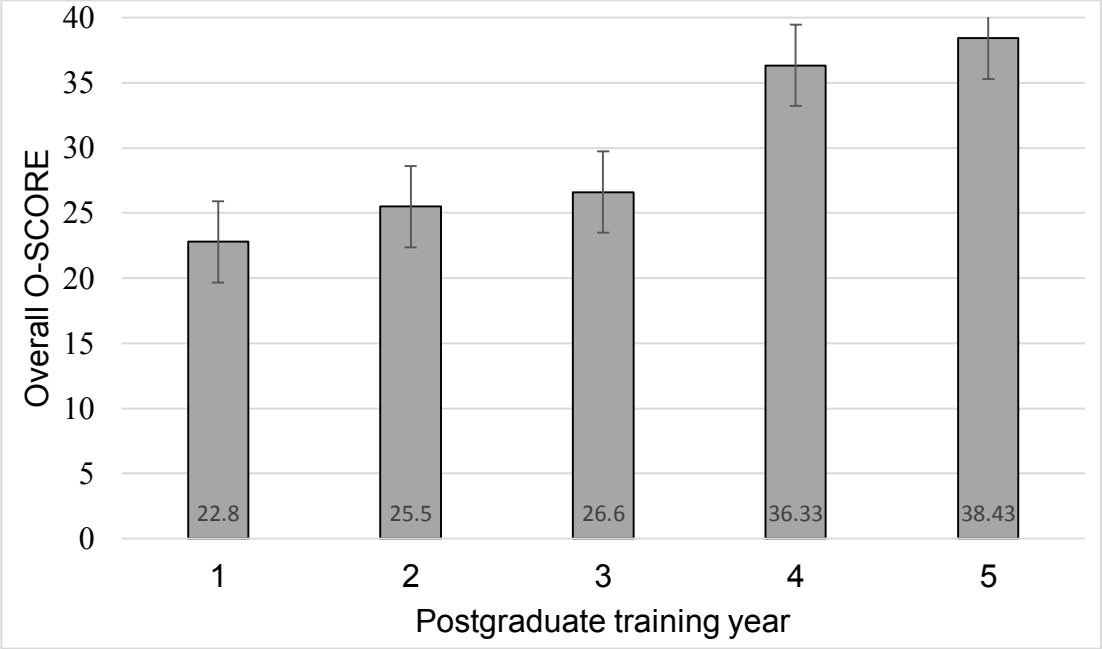

Figure 2- Mean overall O-SCORE results based on postgraduate year of training. The lowest possible O-SCORE is eight and the maximum score is 40.

To further examine the differences a t-test was conducted. The independent variables were junior and senior residents. The dependent variables were all eight items on the O-SCORE which are scored using the entrustability scale and the overall score. In total nine t tests were performed. The normality assumption was met. For junior residents Shapiro-Wilks statistic = 0.94, $\nu = 18$, $p = 0.326$. For senior residents Shapiro-Wilks statistic = 0.82, $\nu = 23$, $p = 0.001$. As this test is sensitive to deviations from normality, a more stringent significance level of $p < 0.001$ was used. These results show that normality assumption was met. All eight items scored with an entrustability scale and the overall scores were significantly different between the two groups. The results are provided in Table 2. Levene's test for homogeneity of variance was used to determine whether assumption for homogeneity of variance was met. In cases where it was not met, the F value assigned to unequal variances was used.

Table 2

Mean O-SCORE item scores and standard deviations for junior and senior Plastic Surgery residents

Item	Junior Resident (PGY 1-3)	Senior Resident (PGY 4-5)	p-value
Preprocedural plan	2.67 ±1.37	4.91 ±0.29	< 0.001
Case preparation	3.22 ±0.94	4.87 ±0.34	< 0.001
Knowledge of steps	2.89 ±0.83	4.70 ±0.47	< 0.001
Technical skills	2.89 ±0.83	4.39 ±0.89	< 0.001
Visuospatial skills	3.17 ±0.78	4.39 ±0.72	< 0.001
Postprocedural plan	3.78 ±0.88	4.33 ±0.39	< 0.001
Efficiency and flow	2.94 ±1.06	4.61 ±0.66	< 0.001
Communication	3.67 ±0.97	4.91 ±0.42	< 0.001
Overall score	25.06 ±5.91	37.61±2.68	< 0.0001

The ninth item on the O-SCORE is a dichotomously score item. It asks evaluators to decide whether the trainee is ready to perform the procedure independently. In this study, a Chi square test was used to determine difference in ratings based on level of training. There was a significant difference in the number of ratings where the trainee was deemed ready to perform procedure independently, $\chi^2 = 29.73$, $\nu = 1$, $p < 0.001$. The differences between groups for the dichotomously scored item is presented in Figure 3.

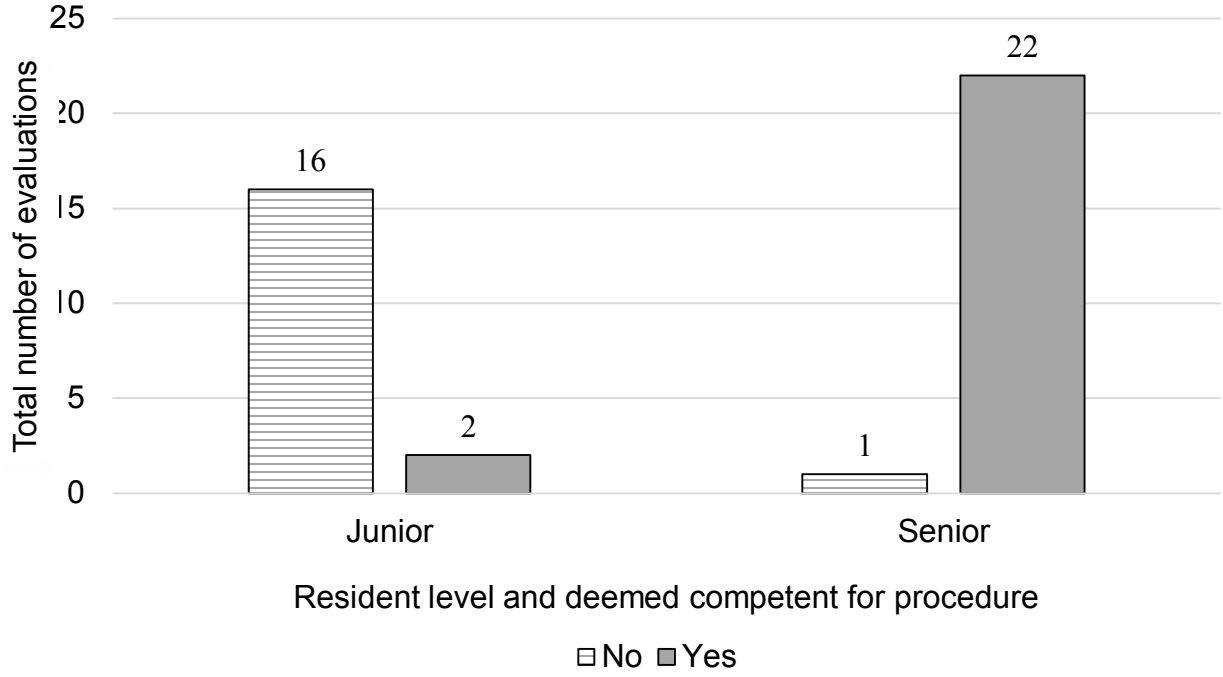

Figure 3- Evaluation of competency on the O-SCORE based on seniority within a Plastic Surgery residency program**.** No indicates the assessor indicated not ready to perform this procedure independently. Yes indicates the assessor indicated the trainee was ready to perform independently.

These initials results show that assessment made using the O-SCORE is able to differentiate between junior and senior residents. There is a statistical difference between the trainees in the first three years of training and those in the final two years. This provides validity evidence for the assessments made using the O-SCORE in Plastic Surgery.

Part 3- Generalizability Study

A generalizability study (G study) was carried out with two facets: occasion and item. The goal of the G study was to determine the sampling variability of the occasions and items of

the O-SCORE when applied to Plastic Surgery residents. Because residents were evaluated on different occasions, occasion was nested within resident. Residents obtained a range of two to six evaluations of three different types. Resident and occasion were treated as random effects. Item as analyzed as a fixed effect. Having a combination of random and fixed facets results in a mixed design and the data was analyzed as such.

In the G study, absolute error was used in lieu of relative error. This absolute error was chosen because the decision to be made using the O-SCORE is based on competence and should not be made relative to other residents in the same training year. The variance components for *r, i, o:r, ri* and *oi:r* were determined to be 0.75, 0.19, 0.04, 0.08 and 0.28 respectively. Of the variables in the generalizability study, residents (55%) contributed most to the variance of the overall score, followed by the error term (21%) and the item variance (14%). These results are summarized in Table 4. The dependability coefficient for the G study was 0.91.

Table 4

Variance components and percent of variance accounted for by each facet

Facet	df	SS	MS	Variance component	Percent %

r	9	236.45	26.27	0.75	55.39
i	7	17.32	2.47	0.19	14.23
o:r	31	56.25	1.81	0.04	3.30
ri	63	38.45	0.61	0.08	6.00
o:ri	217	61.48	0.28	0.28	21.06

r = resident, i = item, o = occasion

Given the high stakes decisions that are to be made in medical education, it is important to maximize the reliability of all assessments. In order to understand the effect of changing assessment parameters a decision study (D study) was conducted.

Figure 4 reflects the outcomes on the dependability coefficient and absolute variance when assessment parameters are adjusted. Increasing the number of occasions increases the overall reliability of the assessment. In this D Study, the item facet was left fixed at eight as the reliability index of the assessment is only changed by the number of assessments completed. For items to be changed, the O-SCORE assessment tool would have to be changed. From the D Study it is evident that at least two occasions must be evaluated in order to increase the reliability to > 0.85.

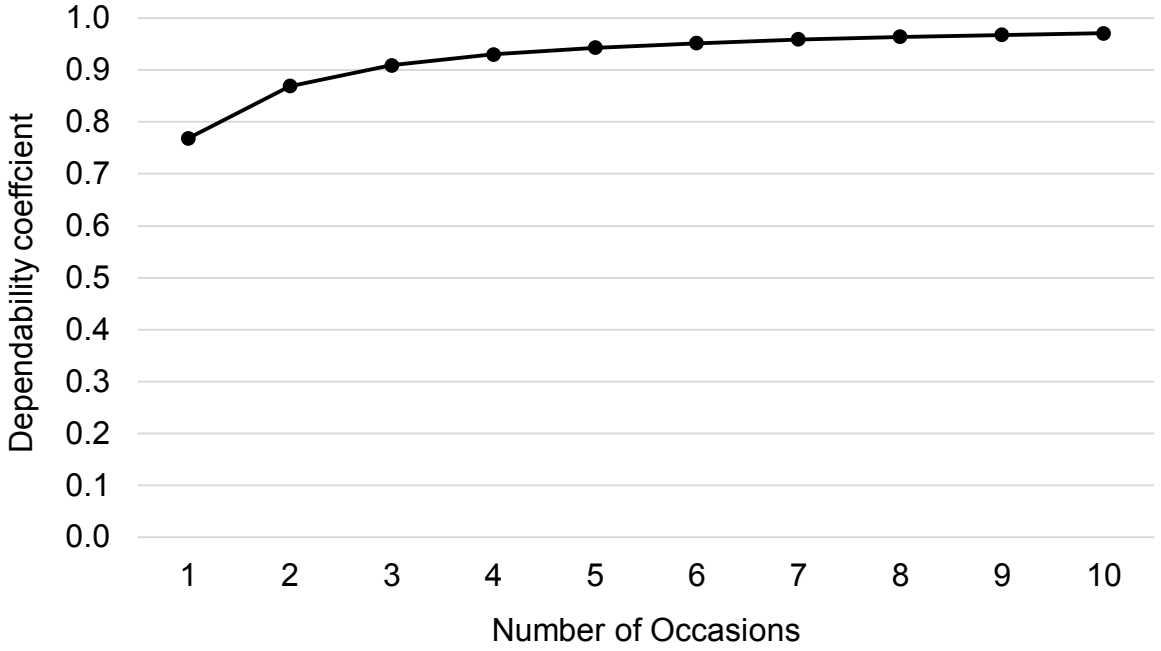

Figure 4- Changes in dependability coefficient when number of occasions using O-SCORE is increased

From the generalizability study it was shown that the O-SCORE offers reliable assessment across items and occasions when used in a Plastic Surgery training program. The dependability coefficient in this study was 0.91. The decision study showed that only two occasions are required to obtain a reliability coefficient > 0.85, providing a feasibility argument for this type of workplace-based assessment.

Discussion

Assessment in competency-based education is not a straightforward task. The purpose of this study was to evaluate the reliability of a workplace-based assessment. It was shown that in a single institution, the O-SCORE can offer reliable assessment results and is able to differentiate between junior and senior residents. Developing an assessment plan or framework to guide training is imperative in the new era of CBME. In addition, understanding the psychometric properties of assessment tools is important to establish validity for assessment use. If not reliable, an assessment does not yield valid test score interpretations (Messick, 1995). This discussion will explore the implications of the findings of this study, describe how the findings compare and contrast to previously published studies, and highlight limitations and future directions.

Differentiation of Trainees

This study has produced numerous positive results which favor using the O-SCORE in a competency-based Plastic Surgery training program. The assessment scores produced by the O-SCORE are reliable, $\Phi = 0.91$. There was a significant difference in item mean score, overall mean score, and the number of times a trainee was deemed competent between PGY 1-3 and PGY 4-5. This differentiation of trainees based on known differences is accepted as a form of construct-validity evidence (Fried & Feldman, 2008) . In this study, there was a significant difference between a junior and senior resident being deemed competent. In the current residency program, trainees are deemed to be senior residents once they enter the third year of training. This shift from junior to senior entails greater on-call responsibility. It is important that competency in all Plastic Surgery domains is not extrapolated from the results of this study. Only hand fracture fixation, mandibular ORIF, and breast reduction mammoplasty were evaluated. A senior resident on call would be responsible for triaging consults, deciding upon treatment plans,

and performing emergency bedside procedures. These skills were not evaluated by the O-SCORE. It may be that the skills necessary to act as on call senior resident are not the same as those needed to be operatively competent on the three surgery types included in this study. The emergency procedures for which competency is necessary may be learned at an early stage in training compared to more complex surgical procedures. Nonetheless, the results of this study show that more senior trainees are deemed to be surgical competent in breast reduction mammoplasty, mandible ORIF and hand fracture fixation more often than are junior trainees. In the original study by Gofton et al. (2012) the scores were significant in three groups (a) PGY 1-2, (b) PGY-3, and (c) PGY-4-5. In this study, the score differences were only significant in two groups (a) PGY 1-3 and (b) PGY 4-5. This difference between studies may be due to sample size, could represent a difference between Orthopedic/ General Surgery training and Plastic Surgery training, and/or represent an institutional difference in expectations. In a Plastic Surgery training program using CORE in the United States, the score distinction was not provided based on year of training. Instead, they reported a 0.4 unit increase in score for each increase in year of training. Regardless, all studies using an entrustability scale show progression as year of training increase. This provides evidence that an entrustability scale does provide validity evidence in surgical training (Cooney et al., 2016a; Gofton et al., 2012).

High Reliability

The reliability of scores from the O-SCORE assessment were quite high in this study. Reliability in the form of internal consistency measures was high for the technical items. It is reassuring that raters are interpreting items measuring the same construct in a similar fashion. The Cronbach alpha for technical items was 0.90. This value is deemed to be excellent on published standards of acceptable reliability parameters (Bland & Altman, 1997). As surgical

training should be taken with utmost responsibility, the assessment measures should be of high quality given the high stakes nature of training. Other tools such as the OSATS, breast augmentation and face lift checklists, and the upper limb evaluation tool have Cronbach alpha values ranging from 0.37 to 0.91. The implication for low reliability is the impact on validity of the assessment. Without being reproducible, it is not possible to deem an assessment valid (Traub & Rowley, 1991). It is clear from this study that internal consistency is very high when using the O-SCORE which ultimately strengthens the validity evidence for its use.

The inter-rater reliability of the O-SCORE when used at the University of Alberta was explored using a blinded video analysis. Standard error of the mean (SEM) was determined based on the assessments of seven raters on three items assessing surgical competence. The SEM ranged from 0.36 to 0.37 for the technical items. The standard deviation for these items ranged from 0.95 to 0.98. This agreement could be considered acceptable. However, if a minimum score based on year of training is to be developed, this range needs to be considered. As residents advance through training they will be expected to perform at a higher level of proficiency. The establishment of an expected progression of skill development was not the focus of this study but will be needed in the future.

The results of the generalizability study were different than reported by Gofton et al. (2012). This study resulted in a much higher reliability coefficient (0.91 versus 0.80). The field of Plastic Surgery is quite broad. The universe of admissible observations however was limited to three surgery types. The study by Gofton et al. (2012) included six procedures in each of orthopedic and general surgery. This was done to include procedures of which junior residents would be acting as assistant or surgeon for a part of the case. Assessment using the O-SCORE did present a challenge for first and second year residents in this thesis study as they were often

accompanied in the operating room by a more senior trainee. This highlights the fact that assessment of surgical skills can be difficult when there is little exposure to the trainee. Here, it is proposed that utilization of the O-SCORE be implemented in a staged fashion. The more basic operative procedures could be introduced to the trainee at an early stage of training and assessments be performed to provide feedback and track progress. As a trainee advances through the program or meets basic cut off criteria, more complex procedures could be introduced to the assessment portfolio. This is a logical approach to evaluation in Plastic Surgery as a first year resident will not be trusted to perform a free flap harvest, for example, and the assessment of such a procedure is futile.

Using the O-SCORE as a Formative Assessment Tool

Although the O-SCORE is not likely to serve as a summative evaluation, the formative assessments which will be used for CBME should be held to a high standard of validity and reliability. A summative evaluation is classically one assessment which can be used to generalize to the target domain with high reliability (Messick, 1995). In the case of a workplace-based performance assessment, multiple assessments of real-life procedures are needed to accurately generalize to the domain of surgical competency making these assessments more appropriate for formative rather than summative evaluation. It has been described that a formative assessment should be "low stakes … and intended to stimulate learning" (Norcini et al., 2011). The rationale for using O-SCORE assessments as a formative assessment tool is multifactorial. These reasons are (a) the progressive nature of training, (b) the need to document progress, and (c) the feedback offered by the O-SCORE tool. These reasons will now be explored. The first reason is because of the progressive nature of surgical training. Trainees are gradually given independence in the operating room. This is due to the complexity of the procedures which are performed, the safety

of the patient, and the rapport of the surgeon with the trainee. The amount of operating a trainee performs is linked to the amount of trust between surgeon and trainee (Rekman et al., 2016). Because this trust is likely to evolve over time, the O-SCORE feedback is useful for the trainee to gauge their progression. Without a progress update, surgical trainees may fail to recognize their shortcomings and need for improvement. This feedback is valuable and repeated O-SCORE measures allow the trainee and program administrators to see the progress being made. One concern is the potential for range restriction in assessment. In this study, all levels of the scale were used when evaluating residents. If range restriction is significant, the feedback to residents may be biased to avoid providing low ratings. This avoidance of true reporting of performance affects validity and reliability. It also affects the quality of the feedback provided. Since the entire range for the O-SCORE scale was used in this study, it is probable that the feedback was honest. This is important for a formative assessment and strengthens the argument for using the O-SCORE as a formative assessment tool.

The second reason for using the O-SCORE as a formative assessment tool is to document the progress of trainees. According to current accreditation standards in Canada, "there must be mechanisms in place to ensure systematic collection and interpretation of assessment data on each resident enrolled in the program" (Royal College of Physicians and Surgeons of Canada, 2011). This standard can be met by implementation of the O-SCORE as a regular assessment of skills. The necessity of documentation of performance is also important for the legal considerations should there be difficulties with a trainee in the future. Any decision where a trainee requires remediation or dismissal will require prior documentation of performance (Irby & Milam, 1989).

The third reason which supports the use of the O-SCORE in formative assessment is the feedback offered by the O-SCORE. The eight scaled items, independence readiness item, and narrative feedback provide ample feedback to trainees. In this study, the item-total correlations for the scaled items are quite high. This may indicate redundancy in the items and that some may not be necessary. Although this may be true, the O-SCORE authors defend this issue by stating that the items offer opportunity of surgeons to lead feedback sessions with trainees (Gofton, Halman, & Wood, 2016).

Subjectivity in Assessment

The O-SCORE has been shown to produce valid and reliable assessment score interpretations despite being a subjective assessment. There has been an argument for objectivity in assessment for many years (Norman et al., 1991). As competency is defined and deconstructed into measureable components, objective assessment seems logical. Whether or not a truly objective assessment can exist in medical education is an important fundamental question. The O-SCORE relies on the judgement of a surgeon to provide an opinion of readiness for independent practice. It is aligned with Miller's pyramid of skills acquisition for the top tier level: "does". Van der Vleuten & Verhoeven (2013) have made an argument that assessing the "does" level should be assessed with enough frequency and from multiple raters to provide an overview of a trainee across many different situations. Using the Ottawa consensus criteria for good assessment, the O-SCORE has shown validity, reproducibility, equivalence, and feasibility. Objective testing of surgical competency is currently very labour intensive and dependent on significant personal and financial resources. One benefit of the O-SCORE is that it is simple to use and offers quick and potentially meaningful feedback to the trainee. The often quoted need for objective assessment is misguided based on previous work in this area. It has been shown that

subjective assessments interpretations can be valid and reliable (Schuwirth LWT. & van der Vleuten CPM., 2012). In an attempt to reduce subjectivity in work-based assessment, checklists have been created for breast augmentation and face lift procedure (Courteau et al., 2015). Although the authors attempt to provide an objective assessment, they risk emphasis on rote memory versus the desired development of surgical skill (Norman et al., 1991). It has been shown that in the hands of an expert a global rating scale, such as that used in the O-SCORE, offers greater validity evidence and inter-station reliability when applied to an observed structured clinical exam (Regehr, MacRae, Reznick, & Szalay, 1998). This research has shown that validity and reliability exist when a subjective assessment is made.

Uniqueness of Study

The necessity for research in the education of Plastic Surgery trainees is high and yet there is little published on training of Plastic Surgeons. This study is Plastic Surgery focused and offers validity evidence for using an entrustability scaled assessment tool. The ACGME milestone project in the United States and Competence by Design in Canada all require increased attention on assessment. Since initiating this study, researchers at John Hopkins University have published their experience using CORE, an entrustability scaled evaluation which rates trainees on one item (Cooney et al., 2016b). This O-SCORE study uses a tool rating trainees using nine items. It also provides results in a format which makes it comparable to other studies. Authors of the CORE study use multiple regression to provide validity evidence but the practicality of the results may be lost in its complexity. Additionally, this thesis study was conducted in Canada and represents the first global rating assessment study in a Canadian Plastic Surgery program. Although few published works describe assessment differences based on geographical area, there exists a definite difference in terms of expectations of trainees, surgical training, and level of

independence in the operating room (Kamali et al., 2016; Norcini, Anderson, & McKinley, 2006).

This study strengthens the support for subjective assessment in surgical training. The opinions of experts in surgery should not be minimized in a pursuit of objectivity. Much of what is deemed exceptional in surgery is not measurable without the use of subjective scales. Fairness and equality in assessment can be assured by having a large number for raters and multiple occasions for assessment. The O-SCORE uses the opinion of experts to decide whether a trainee is ready for independent practice. The correlation of overall O-SCORE with rating of competency based on year of training suggests there is validity in the decisions which are being made. The argument for using the O-SCORE is based on the entrustability scale which is only present in two other assessment tools, Zwisch scale and CORE. The entrustability scale is the unique aspect of the O-SCORE which motivated its use in this study.

Limitations

This study was not without limitations. The first was the small sample size. As a single center study there were limited participants available to participate. The length of time required to collect this data reflects difficulties in performing assessment in a culture where regular performance-based assessments are not the norm. The statistical tests did not seem to be affected by the sample size. It has been shown that reliability coefficients in a generalizability study are most stable when sample size is 400 (Altigan, 2013). It was recommended in the same study that a sample size of 50-300 be used for a generalizability study. More evaluations for this study could have been collected if the type of surgery was not restricted and if an increased data collection period was utilized. In order to maintain some standardization for the study the type of surgery to be evaluated was restricted to three common Plastic Surgery procedures. This

restriction in surgery type ultimately narrows the universe of generalization and decreases reliability estimates.

Another limitation of the study is the length of time over which data was collected. In this study eight months of data collection occurred. It is likely that residents improved in their abilities over this time. There is no measure of the stability of performance assessment over time. Given that trainees were classified by training year and that these groups remained the same throughout the study, this change in ability may be balanced amongst all those in the study. Given the narrow range in scores for trainees at different levels this is likely true.

It has been shown that there are six factors which impact the evaluation of a resident's performance (Williams, Dunnington, & Klamen, 2005).These factors are (a) incomplete sampling of performance, (b) rater memory constraints or distortions, (c) hidden performance deficits of the resident, (d) lack of meaningful benchmarks, (e) faculty members' hesitancy to act on negative performance information, and (f) systematic rater error (Williams et al., 2005). This thesis study did not explore the implications for a series of negative performance ratings. The consequences for this forms part of validity evidence and could be studied in the future.

Limitations of the O-SCORE.

No assessment method is without flaw. The authors of the O-SCORE acknowledge the redundancy in the items however justify the item inclusion as a method to deliver feedback to trainees (Gofton et al., 2016). In this study the item-total correlations are high which further supports the redundancy of the items of the O-SCORE. On further analysis of the items it can be argued that each item measures multiple skills. The items on the O-SCORE measure multiple skills which constitute competency. An example of this is the third O-SCORE item "knowledge of specific procedural steps". The descriptor for this item lists three areas for assessment. These

three abilities are different and yet are assessed using one measure, the score for knowledge of specific procedural steps. This does highlight a flaw in the item design of the O-SCORE. This flaw is that individual items are assessing multiple constructs. Interestingly however, there is a high internal consistency of items and high reliability when multiple raters assign scores to resident performance using the O-SCORE. This is interesting because the multiple skills assessed using each item is being interpreted with a high degree of consistency between raters. Ghaderi et al (2014) compared multiple surgical assessment tools in terms of validity evidence. The O-SCORE was shown to produce considerable evidence as per current criteria of evidence for construct validity. This evidence was strongest in internal consistency, as the original O-SCORE authors conducted a generalizability study and performed item psychometric analysis.

Future Directions

There are three directions for future research. The first is to implement the O-SCORE as a formative assessment tool in multiple training centers. Evaluating the reliability of the tool at multiple sites will allow a comparison of geography and assessment culture across the country using mean scores and psychometric analyses. A similar generalizability study to this thesis study could be conducted at multiple sites thereby permitting score comparisons. It has been shown in other studies that variability in scores increases when assessments are conducted in multiple centers (Gao & Others, 1994). This information will be valuable to the Royal College of Physicians and Surgeons of Canada as national standards of assessment may be required as CBME is implemented. It would also be the first multi-university assessment study in Plastic Surgery.

The second direction for future research is to examine O-SCORE scores for trainees longitudinally. The relevance of this research is to determine how residents progress and will

allow the determination of minimal acceptable scores for training levels. This is important for CBME as a duration of training cut off will be required to allow for logistics of organizing programs. One objective of CBME is the early identification of struggling residents and establishing a progression model will help administrators identify these trainees.

Third, the results of the O-SCORE evaluations need to be compared to other evaluations. This comparison may be conducted using other surgical evaluation tools. Additionally, research is required to understand how performance assessment fits into the overall assessment of trainees. There has been some prior work examining resident portfolios (O'Sullivan, Reckase, McClain, Savidge, & Clardy, 2004; Roberts, Shadbolt, Clark, & Simpson, 2014). The usefulness of the O-SCORE in these portfolios should be explored. This may be conducted by comparing scores from the O-SCORE and outcomes such as pass rates of licensing exams as well as other evaluation tools assessing different constructs. The correlations of technical and non-technical constructs need to be explored further.

Conclusion

The O-SCORE and its entrustability-scaled anchors offer a construct-aligned assessment tool for operative skills. At the University of Alberta's Plastic Surgery residency program, it has been shown that this tool offers high internal consistency of technical items, high overall reliability with three occasions, and validity evidence. Differentiation of trainees based on year of training is possible using the O-SCORE. The utility of this tool seems to be in formative assessment. Although other tools are available for Plastic Surgery programs to use, none offer entrustability-scaled anchors and have provided psychometric evidence of high reliability and validity as has the O-SCORE tool assessment. As many countries shift from an apprenticeship model of training to CBME, there will be a need for frequent assessments and documented

assessment outcomes. The O-SCORE can offer important feedback to trainees and programs about the readiness of trainees to operate independently. Assuring competence is the goal of CMBE and using the O-SCORE as part of an assessment armamentarium is a step closer to this objective.

References

Alam, M., Nodzenski, M., Yoo, S., Poon, E., & Bolotin, D. (2014). Objective structured assessment of technical skills in elliptical excision repair of senior dermatology residents: A multirater, blinded study of operating room video recordings. *JAMA Dermatology, 150*(6), 608-612. doi:10.1001/jamadermatol.2013.6858 [doi]

Altigan, H. (2013). Sample size for estimation of G and phi coefficients in generalizability theory. *Eurasian Journal of Educational Research, Spring*(51), 215-228.

Bland, J. M., & Altman, D. G. (1997). Cronbach's alpha. *BMJ (Clinical Research Ed.), 314*(7080), 572.

Bloch, R., & Norman, G. (2011). *G string IV* (6.1.1 ed.). Hamilton, ON, Canada: Ralph Bloch & Geoff Norman.

Bodle, J. F., Kaufmann, S. J., Bisson, D., Nathanson, B., & Binney, D. M. (2008). Value and face validity of objective structured assessment of technical skills (OSATS) for work based assessment of surgical skills in obstetrics and gynaecology. *Medical Teacher, 30*(2), 212-216. doi:10.1080/01421590701881624 [doi]

Brennan, R. (1992). NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practice, Winter*, 27-34.

Brennan, R. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353.

Carmine, E., & Zeller, R. (1979). In Sullivan J. (Ed.), *Reliability and validity assessment*. United States: Sage Publications.

Carraccio, C., Wolfsthal, S. D., Englander, R., Ferentz, K., & Martin, C. (2002). Shifting paradigms: From flexner to competencies. *Academic Medicine : Journal of the Association of American Medical Colleges, 77*(5), 361-367.

Carraccio, C. L., & Englander, R. (2013). From flexner to competencies: Reflections on a decade and the journey ahead. *Academic Medicine : Journal of the Association of American Medical Colleges, 88*(8), 1067-1073. doi:10.1097/ACM.0b013e318299396f [doi]

Chivers, Q. J., Ahmad, J., Lista, F., Warren, R. J., Arkoubi, A. Y., Mahabir, R. C., . . . Islur, A. (2013). Cosmetic surgery training in canadian plastic surgery residencies: Are we training competent surgeons? *Aesthetic Surgery Journal / the American Society for Aesthetic Plastic Surgery, 33*(1), 160-165. doi:10.1177/1090820X12467794 [doi]

Cizek, G. J. (1991). Innovation or enervation? performance assessment in perspective. *Phi Delta Kappan, 72*(9), 695-99. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ425523&site=ehost-live&scope=site

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cooney, C. M., Cooney, D. S., Bello, R. J., Bojovic, B., Redett, R. J., & Lifchez, S. D. (2016a). Comprehensive observations of resident evolution: A novel method for assessing procedure-

based residency training. *Plastic and Reconstructive Surgery, 137*(2), 673-678. doi:10.1097/01.prs.0000475797.69478.0e [doi]

Cooney, C. M., Cooney, D. S., Bello, R. J., Bojovic, B., Redett, R. J., & Lifchez, S. D. (2016b). Comprehensive observations of resident evolution: A novel method for assessing procedure-based residency training. *Plastic and Reconstructive Surgery, 137*(2), 673-678. doi:10.1097/01.prs.0000475797.69478.0e [doi]

Cooney, C. M., Redett, R. J.,3rd, Dorafshar, A. H., Zarrabi, B., & Lifchez, S. D. (2014). Integrating the NAS milestones and handheld technology to improve residency training and assessment. *Journal of Surgical Education, 71*(1), 39-42. doi:10.1016/j.jsurg.2013.09.019 [doi]

Cooper, H. M., & American Psychological Association. (2012). *APA handbook of research methods in psychology* (1st ed.). Washington, D.C.: American Psychological Association. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&NEWS=N&PAGE=toc&SEARCH=2011-23865.dd&LINKTYPE=asBody&D=psbk; http://search.ebscohost.com/direct.asp?db=pzh&jid=%22201123865%22&scope=site; http://search.ebscohost.com/login.aspx?direct=true&db=pzh&jid=201123865&site=ehost-live

Courteau, B. C., Knox, A. D., Vassiliou, M. C., Warren, R. J., & Gilardino, M. S. (2015). The development of assessment tools for plastic surgery competencies. *Aesthetic Surgery*

Journal / the American Society for Aesthetic Plastic Surgery, 35(5), 611-617. doi:10.1093/asj/sju068 [doi]

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=psyc1&AN=1952-03137-001; http://resolver.library.ualberta.ca/resolver?sid=OVID:psycdb&id=pmid:&id=doi:10.1007%2FBF02310555&issn=0033-3123&isbn=&volume=16&issue=3&spage=297&pages=297-334&date=1951&title=Psychometrika&atitle=Coefficient+alpha+and+the+internal+structure+of+tests.&aulast=Cronbach&pid=%3Cauthor%3ECronbach%2C+Lee+J%3C%2Fauthor%3E%3CAN%3E1952-03137-001%3C%2FAN%3E%3CDT%3EJournal+Article%3C%2FDT%3E

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ690014&site=ehost-live&scope=site; http://dx.doi.org/10.1177/0013164404266386

Crossley, J., Humphris, G., & Jolly, B. (2002). Assessing health professionals. *Medical Education, 36*(9), 800-804. doi:1294 [pii]

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Medical Education, 45*(6), 560-569. doi:10.1111/j.1365-2923.2010.03913.x [doi]

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. *Medical Education, 46*(1), 28-37. doi:10.1111/j.1365-2923.2011.04166.x [doi]

DaRosa, D. A., Zwischenberger, J. B., Meyerson, S. L., George, B. C., Teitelbaum, E. N., Soper, N. J., & Fryer, J. P. (2013). A theory-based model for teaching and assessing residents in the operating room. *Journal of Surgical Education, 70*(1), 24-30. doi:10.1016/j.jsurg.2012.07.007 [doi]

Downing, S. M. (2003). Validity: On meaningful interpretation of assessment data. *Medical Education, 37*(9), 830-837.

Doyle, J. D., Webber, E. M., & Sidhu, R. S. (2007). A universal global rating scale for the evaluation of technical skills in the operating room. *American Journal of Surgery, 193*(5), 551-5; discussion 555. doi:S0002-9610(07)00074-8 [pii]

Dumestre, D., Yeung, J. K., & Temple-Oberle, C. (2014). Evidence-based microsurgical skill-acquisition series part 1: Validated microsurgical models--a systematic review. *Journal of Surgical Education, 71*(3), 329-338. doi:10.1016/j.jsurg.2013.09.008 [doi]

Ferron, C. E., Lemaine, V., Leblanc, B., Nikolis, A., & Brutus, J. P. (2010). Recent canadian plastic surgery graduates: Are they prepared for the real world? *Plastic and Reconstructive Surgery, 125*(3), 1031-1036. doi:10.1097/PRS.0b013e3181cb6128 [doi]

Flexner, A. (1910). *Medical education in the united states and canada; a report to the carnegie foundation for the advancement of teaching*. New York: Carnegie Foundation for the Advancement of Teaching.

Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., De Rossi, S., & Horsley, T. (2010). Toward a definition of competency-based education in medicine: A systematic review of published definitions. *Medical Teacher, 32*(8), 631-637. doi:10.3109/0142159X.2010.500898 [doi]

Frank, J. R., Snell, L. S., Cate, O. T., Holmboe, E. S., Carraccio, C., Swing, S. R., . . . Harris, K. A. (2010). Competency-based medical education: Theory to practice. *Medical Teacher, 32*(8), 638-645. doi:10.3109/0142159X.2010.501190 [doi]

Fried, G. M., & Feldman, L. S. (2008). Objective assessment of technical performance. *World Journal of Surgery, 32*(2), 156-160. doi:10.1007/s00268-007-9143-y [doi]

Gao, X., & Harris, D. (2012). Generalizability theory. In H. Cooper (Ed.), *APA handbook of research methods in psychology, volume 1* (pp. 661-681). United States: American Psychological Association. doi:10.1037/13619-035

Gao, X., & Others, A. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*(4), 323-42. Retrieved from

http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ495741&site=ehost-live&scope=site

George, B. C., Teitelbaum, E. N., Meyerson, S. L., Schuller, M. C., DaRosa, D. A., Petrusa, E. R., . . . Fryer, J. P. (2014). Reliability, validity, and feasibility of the zwisch scale for the assessment of intraoperative performance. *Journal of Surgical Education, 71*(6), e90-6. doi:10.1016/j.jsurg.2014.06.018 [doi]

Ghaderi, I., Manji, F., Park, Y. S., Juul, D., Ott, M., Harris, I., & Farrell, T. M. (2015). Technical skills assessment toolbox: A review using the unitary framework of validity. *Annals of Surgery, 261*(2), 251-262. doi:10.1097/SLA.0000000000000520 [doi]

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. *Academic Medicine : Journal of the Association of American Medical Colleges, 85*(5), 780-786. doi:10.1097/ACM.0b013e3181d73fb6 [doi]

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social & Administrative Pharmacy : RSAP, 9*(3), 330-338. doi:10.1016/j.sapharm.2012.04.004 [doi]

Gofton, W. T., Dudek, N. L., Wood, T. J., Balaa, F., & Hamstra, S. J. (2012). The ottawa surgical competency operating room evaluation (O-SCORE): A tool to assess surgical competence. *Academic Medicine : Journal of the Association of American Medical Colleges, 87*(10), 1401-1407. doi:10.1097/ACM.0b013e3182677805 [doi]

Gofton, W. T., Halman, S., & Wood, T. (2016). Competency based assessment tools using "entrustability anchors". *Ottawa Conference,* Perth, Australia.

Grant, G. (1979). *On competence : A critical analysis of competence-based reforms in higher education* San Francisco : Jossey-Bass Publishers, 1979; 1st ed. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat03710a&AN=alb.248040&site=eds-live&scope=site

Haertel, E. (2006). Reliability. In R. Brennan (Ed.), *Educational measurement* (Fourth ed., pp. 65-110). United States: Praeger Publishers.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Irby, D. M., & Milam, S. (1989). The legal context for evaluating and dismissing medical students and residents. *Academic Medicine : Journal of the Association of American Medical Colleges, 64*(11), 639-643.

Kamali, P., van Paridon, M., Ibrahim, A., Paul, M., Winters, H., Martinnot-Duquennoy, V., . . . Lin, S. (2016). Plastic surgery training worldwide: Part 1. the united states and europe. *PRS Global Open, 4*(3), e641. doi:10.1097/GOX.0000000000000627

Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*(2), 125-160.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, Summer*, 5-17.

Kassam, A., Donnon, T., & Rigby, I. (2014). Validity and reliability of an in-training evaluation report to measure the CanMEDS roles in emergency medicine residents. *Cjem, 16*(2), 144-150.

Kasten, S. J., Levi, B., Eng, D., & Schenarts, K. A. (2009). Toward outcomes-based plastic surgery training: A needs assessment of recent graduates. *Plastic and Reconstructive Surgery, 124*(5), 1703-1710. doi:10.1097/PRS.0b013e3181b98c49 [doi]

Kennedy, T. J., Regehr, G., Baker, G. R., & Lingard, L. (2008). Point-of-care assessment of medical trainee competence for independent clinical work. *Academic Medicine : Journal of the Association of American Medical Colleges, 83*(10 Suppl), S89-92. doi:10.1097/ACM.0b013e318183c8b7 [doi]

Knox, A. D., Gilardino, M. S., Kasten, S. J., Warren, R. J., & Anastakis, D. J. (2014). Competency-based medical education for plastic surgery: Where do we begin? *Plastic and Reconstructive Surgery, 133*(5), 702e-710e. doi:10.1097/PRS.0000000000000082 [doi]

MacEwan, M. J., Dudek, N. L., Wood, T. J., & Gofton, W. T. (2016). Continued validation of the O-SCORE (ottawa surgical competency operating room evaluation): Use in the simulated environment. *Teaching and Learning in Medicine, 28*(1), 72-79. doi:10.1080/10401334.2015.1107483 [doi]

Martin, J. A., Regehr, G., Reznick, R., MacRae, H., Murnaghan, J., Hutchison, C., & Brown, M. (1997). Objective structured assessment of technical skill (OSATS) for surgical residents. *The British Journal of Surgery, 84*(2), 273-278.

McGathie, W., Miller, G., Sajid, A. & Telder, T. (1978). Competency-based curriculum development in medical education. Retrieved from http://apps.who.int/iris/bitstream/10665/39703/1/WHO_PHP_68.pdf

McGoey, K. E., Cowan, R. J., Rumrill, P. P., & LaVogue, C. (2010). Understanding the psychometric properties of reliability and validity in assessment. *Work, 36*(1), 105-111 7p. doi:10.3233/WOR-2010-1012

McGrath, M. H. (2014). The plastic surgery milestone project. *Journal of Graduate Medical Education, 6*(1 Suppl 1), 222-224. doi:10.4300/JGME-06-01s1-25 [doi]

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, Winter*, 5-8.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ482658&site=ehost-live&scope=site

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine : Journal of the Association of American Medical Colleges, 65*(9 Suppl), S63-7.

Niitsu, H., Hirabayashi, N., Yoshimitsu, M., Mimura, T., Taomoto, J., Sugiyama, Y., . . . Takiyama, W. (2013). Using the objective structured assessment of technical skills

(OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surgery Today, 43*(3), 271-275. doi:10.1007/s00595-012-0313-7 [doi]

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., . . . Roberts, T. (2011). Criteria for good assessment: Consensus statement and recommendations from the ottawa 2010 conference. *Medical Teacher, 33*(3), 206-214. doi:10.3109/0142159X.2011.551559 [doi]

Norcini, J., Anderson, M. B., & McKinley, D. W. (2006). The medical education of united states citizens who train abroad. *Surgery, 140*(3), 338-346. doi:S0039-6060(06)00322-9 [pii]

Norman, G. R., Van der Vleuten, C. P., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Medical Education, 25*(2), 119-126.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*(1), 1-13.

O'Sullivan, P. S., Reckase, M. D., McClain, T., Savidge, M. A., & Clardy, J. A. (2004). Demonstration of portfolios to assess competency of residents. *Advances in Health Sciences Education : Theory and Practice, 9*(4), 309-323. doi:5270885 [pii]

Plastic Surgery Milestone Working Group.The plastic surgery milestone project: Assessment tools. Retrieved from https://www.med.unc.edu/plastic/education/acgme-documents/milestone-assessment-tools

Pollock, R. A., Donnelly, M. B., Plymale, M. A., Stewart, D. H., & Vasconez, H. C. (2008). 360-degree evaluations of plastic surgery resident accreditation council for graduate medical education competencies: Experience using a short form. *Plastic and Reconstructive Surgery, 122*(2), 639-649. doi:10.1097/PRS.0b013e31817d5fbd [doi]

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine : Journal of the Association of American Medical Colleges, 73*(9), 993-997.

Rekman, J., Gofton, W., Dudek, N., Gofton, T., & Hamstra, S. J. (2016). Entrustability scales: Outlining their usefulness for competency-based clinical assessment. *Academic Medicine : Journal of the Association of American Medical Colleges, 91*(2), 186-190. doi:10.1097/ACM.0000000000001045 [doi]

Ricketts, C. (2009). A plea for the proper use of criterion-referenced tests in medical assessment. *Medical Education, 43*(12), 1141-1146. doi:10.1111/j.1365-2923.2009.03541.x [doi]

Roberts, C., Shadbolt, N., Clark, T., & Simpson, P. (2014). The reliability and validity of a portfolio designed as a programmatic assessment of performance in an integrated clinical placement. *BMC Medical Education, 14*, 197-6920-14-197. doi:10.1186/1472-6920-14-197 [doi]

Royal College of Physicians and Surgeons of Canada. (2011). *General standards applicable to all residency programs.* (). Canada: Royal College of Physicians and Surgeons of Canada.

Royal College of Physicians and Surgeons of Canada. (2014). *Competence by design: Reshaping canadian medical education.* (No. March).

Schuwirth LWT., & van der Vleuten CPM. (2012). Assessing competence: Extending the approaches to reliability. In B. Hodges, & L. Lingard (Eds.), *The question of competence* (pp. 113-130). United States: Cornell University Press.

Shalhoub, J., Vesey, A. T., & Fitzgerald, J. E. (2014). What evidence is there for the use of workplace-based assessment in surgical training? *Journal of Surgical Education, 71*(6), 906-915. doi:10.1016/j.jsurg.2014.03.013 [doi]

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, Calif.: Sage Publications.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations:Uses in assessing rater reliability. *Psychological Bullentin, 86*(2), 420-428.

Spady, W. (1977). Competency based education: A bandwagon in search of a definition. *Educational Researcher, 6*(1), 9-14.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72-101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Standards for educational and psychological testing (1999). Washington, DC American Educational Research Association, c1999. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat03710a&AN=alb.2476536&site=eds-live&scope=site

Streiner, D. L., Norman, G. R., & Cairney, J. (2014). *Health measurement scales: A practical guide to their development and use* (Fifth ed.). Oxford: Oxford University Press. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=891138

Swanson, D. B., & Others, A. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24*(5), 5-11,35. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ509382&site=ehost-live&scope=site

Tavakol, M., & Dennick, R. (2011). Making sense of cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. doi:10.5116/ijme.4dfb.8dfd

ten Cate, O. (2005). Entrustability of professional activities and competency-based training. *Medical Education, 39*(12), 1176-1177. doi:MED2341 [pii]

ten Cate, O., & Scheele, F. (2007a). Competency-based postgraduate training: Can we bridge the gap between theory and clinical practice? *Academic Medicine : Journal of the Association of American Medical Colleges, 82*(6), 542-547. doi:10.1097/ACM.0b013e31805559c7 [doi]

ten Cate, O., & Scheele, F. (2007b). Competency-based postgraduate training: Can we bridge the gap between theory and clinical practice? *Academic Medicine : Journal of the Association of American Medical Colleges, 82*(6), 542-547. doi:10.1097/ACM.0b013e31805559c7 [doi]

Traub, R., & Rowley, G. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, Spring*, 37-45.

Turnbull, J. M. (1989). What is ... normative versus criterion-referenced assessment. *Medical Teacher, 11*(2), 145-150.

van der Vleuten, C., & Verhoeven, B. (2013). In-training assessment developments in postgraduate education in europe. *ANZ Journal of Surgery, 83*(6), 454-459. doi:10.1111/ans.12190 [doi]

Van Heest, A., Putnam, M., Agel, J., Shanedling, J., McPherson, S., & Schmitz, C. (2009). Assessment of technical skills of orthopaedic surgery residents performing open carpal tunnel release surgery. *The Journal of Bone and Joint Surgery.American Volume, 91*(12), 2811-2817. doi:10.2106/JBJS.I.00024 [doi]

VanHeest, A., Kuzel, B., Agel, J., Putnam, M., Kalliainen, L., & Fletcher, J. (2012). Objective structured assessment of technical skill in upper extremity surgery. *The Journal of Hand Surgery, 37*(2), 332-7. 337.e1-4. doi:10.1016/j.jhsa.2011.10.050 [doi]

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of Statistics, 26*, 81-124. doi:10.1016/S0169-7161(06)26004-8

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-13. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ388723&site=ehost-live&scope=site

Williams, R. G., Dunnington, G. L., & Klamen, D. L. (2005). Forecasting residents' performance--partly cloudy. *Academic Medicine : Journal of the Association of American Medical Colleges, 80*(5), 415-422. doi:80/5/415 [pii]

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*(4), 270-292. doi:10.1207/S15328015TLM1504_11 [doi]

Appendix

The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE)

Trainee #:	Level: 1 2 3 4 5	Staff:
Procedure:		Date:

Relative complexity of this procedure to average of same procedure Low Medium High

The purpose of this scale is to evaluate the trainee's ability to perform this procedure safely and independently. With that in mind please use the scale below to evaluate each item, irrespective of the resident's level of training in regards to *this* case.

Scale

1—"I had to do"—*i.e., Requires complete hands on guidance, did not do, or was not given the opportunity to do*

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

3—"I had to prompt them from time to time"—*i.e., Demonstrates some independence, but requires intermittent direction*

4—"I needed to be in the room just in case"—*i.e., Independence but unaware of risks and still requires supervision for safe practice*

5—"I did not need to be there"—*i.e., Complete independence, understands risks and performs safely, practice ready*

1. Preprocedure plan Gathers/assesses required information to reach diagnosis and determine correct procedure required	**1**	**2**	**3**	**4**	**5**
2. Case preparation Patient correctly prepared and positioned, understands approach and required instruments, prepared to deal with probable complications	**1**	**2**	**3**	**4**	**5**
3. Knowledge of specific procedural steps Understands steps of procedure, potential risks, and means to avoid/overcome them	**1**	**2**	**3**	**4**	**5**
4. Technical performance Efficiently performs steps, avoiding pitfalls and respecting soft tissues	**1**	**2**	**3**	**4**	**5**
5. Visuospatial skills 3D spatial orientation and able to position instruments/hardware where intended	**1**	**2**	**3**	**4**	**5**
6. Postprocedure plan Appropriate complete post procedure plan	**1**	**2**	**3**	**4**	**5**
7. Efficiency and flow Obvious planned course of procedure with economy of movement and flow	**1**	**2**	**3**	**4**	**5**
8. Communication Professional and effective communication/utilization of staff	**1**	**2**	**3**	**4**	**5**
9. Resident is able to safely perform *this* procedure *independently* (circle)		**Y**		**N**	
10. Give at least 1 *specific* aspect of procedure done well					
11. Give at least 1 *specific* suggestion for improvement					

Signatures: Staff:

Trainee:

Permission

WOLTERS KLUWER HEALTH, INC. LICENSE TERMS AND CONDITIONS

May 03, 2016

This Agreement between Curtis R Budden ("You") and Wolters Kluwer Health, Inc. ("Wolters Kluwer Health, Inc.") consists of your license details and the terms and conditions provided by Wolters Kluwer Health, Inc. and Copyright Clearance Center.

License Number	3861530042023
License date	May 03, 2016
Licensed Content Publisher	Wolters Kluwer Health, Inc.
Licensed Content Publication	Academic Medicine
Licensed Content Title	The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): A Tool to Assess Surgical Competence.
Licensed Content Author	Gofton, Wade; MD, MEd; Dudek, Nancy; MD, MEd; Wood, Timothy; Balaa, Fady; MD, MEd; Hamstra, Stanley
Licensed Content Date	Jan 1, 2012
Licensed Content Volume Number	87
Licensed Content Issue Number	10
Type of Use	Dissertation/Thesis
Requestor type	Individual
Portion	Figures/table/illustration
Number of figures/tables/illustrations	1
Figures/tables/illustrations used	Appendix 1- The Ottawa Surgical Competency Operative Room Evaluation (O-SCORE)
Author of this Wolters Kluwer article	No
Title of your thesis / dissertation	Using the Ottawa Surgical Competency Operative Room Evaluation (O-SCORE) in a Canadian Plastic Surgery Program
Expected completion date	Jun 2016
Estimated size(pages)	83
Requestor Location	Curtis R Budden 310-10531 117 St Edmonton, AB T5H0A8 Canada Attn: Curtis R Budden
Billing Type	Invoice
Billing Address	Curtis R Budden 310-10531 117 St