Improving Evaluation of the CanMEDS Collaborator Role: Reliability of the Interprofessional Collaborator Assessment Rubric (ICAR) and Gender Bias in Multi-Source Feedback

by

© Mark Hayward B.Sc. B.Ed.

A thesis submitted to the

School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Masters of Science

Faculty of Medicine

Memorial University of Newfoundland

May 2014

St. John's Newfoundland

Abstract

Since the inception of the Royal College of Physicians and Surgeons of Canada (RCPSC) CanMEDS framework, there has been inequality between the assessment of the Medical Expert role and the six non-Medical Expert roles. The purpose of the study was to evaluate the reliability of the use of the Interprofessional Collaborator Assessment Rubric (ICAR) in a multi-source feedback (MSF) approach for assessing post-graduate medical residents' CanMEDS Collaborator competencies. A secondary investigation attempted to determine whether characteristics of raters (i.e., experience, gender, or frequency of interaction with resident) had any influence on overall ICAR score. The ICAR is a 17-item (and global score) assessment tool utilizing a 9-point scale and two open-text responses. The study involved medical residents receiving ICAR assessments from three (3) rater groups (physicians, nurses, and allied health professionals) over a single four-week rotation. Residents were recruited from four (4) unique medical disciplines. Of those participating residents, sixteen (16) residents were randomly chosen. Six (6) of those received at least two (2) assessments from each rater group and were included in the analysis. All nurses and allied health professionals in participating medical / surgical units were invited to participate and were excluded from analysis if they were absent for at least one week of normal shift work or explicitly stated they did not interact with resident. Physicians were self-appointed by the residents. Statistical analysis utilized Cronbach's alpha, compared overall ICAR scores using one-way and two-way, repeated

measures ANOVA, and logistic regression. Missing data using a single imputation stochastic regression method and was compared to the missing data from a pilot study using pair-sample t-test. Results revealed a high response rate (76.2%) with a statistically significant difference between the gender distributions in each rater group, male physicians (81.8%), female nurses (92.5%), and female allied health professionals (88.4%), $p < .001$. Missing data decreased from 13.1% using daily assessments to 8.8% utilizing an MSF process, $p = .032$. An overall Cronbach's alpha coefficient of $\alpha = .981$ revealed high internal consistency reliability. Each ICAR domain also demonstrated high internal consistency, ranging between .881 - .963. The profession of the rater yielded no significant effect with a very small effect size ($F_{2,5} = 1.225$, $p = .297$, $\eta^2 = .016$). The only significant, main-effect on overall ICAR score was found to the gender of the rater ($F_{1,5} = 7.184$, $p = .008$, $\eta^2 = .045$). Female raters scored residents significantly lower than male raters (6.12 v. 6.82). Logistic regression analysis revealed that male raters were 3.08 times more likely than female raters to provide an overall ICAR score of above 6.0 ($p = .013$) and 3.28 times more likely to score above 7.0 ($p = .005$). A significant interaction effect resulted from a two-way repeated measures ANOVA analysis involving the frequency of interaction between raters and residents across items ($F = 2.103$, $p = .025$, $\eta^2 = .014$). The study findings suggest that the use of the modified ICAR form in a MSF assessment process could be a feasible assessment approach to providing formative feedback to post-graduate medical residents on Collaborator competencies.

Acknowledgements

I would like to start by thanking Dr. John Harnett for his continual encouragement and support of my academic endeavours. He always kept an ear to the ground and when the situation presented itself he was grateful enough to introduce me to my future supervisory committee.

This specific thesis would not have been possible without the tremendous research previous conducted by Dr. Vernon Curran and his colleagues. His ever-increasing knowledge of, and interest in, interprofessional education and collaboration made my life so much easier during these two years. There was never a moment when I was without a reference or a resource. I thank you for being so readily available to address my innumerable questions.

Although Dr. Sean Murphy may be a man of few words, he was full of encouraging, positive advice whenever I needed guidance in any aspect of my thesis (statistics, organization, or grammar). He also must be acknowledged for his support of other research projects I envisioned. I look forward to collaborating with him on those in the future.

I am forever in debt to Dr. Henry Schultz for his statistical wizardry. I would not have been able to complete my thesis without his passion for statistics and his willingness to help students, such as myself, who demonstrate interest in understanding and appreciating the statistical methods at play in their research. I hope I learn one tenth of his knowledge in my lifetime.

I must send my most sincere gratitude to my primary supervisor, Dr. Bryan Curtis. He far surpassed his role as a supervisor by also becoming my academic advisor, frequent life coach, colleague, mentor, and a friend. A thank you is not enough for the role you have played in the past two years of my academic life. I look forward to future research together as well as learning from you in the medical world.

I would also wish to thank all individuals who participated in my study, specifically division managers who helped co-ordinate schedules and meeting times with staff.

Finally, to incredible partner Kristy, my mother, father, and brother for putting up with my incessant questions or ramblings about my thesis. Without your support, I'm not sure if this would have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAMC	Association of American Medical Colleges
ACGME	Accreditation Council of Graduate Medical Education
ACS	American College of Surgeons
ANOVA	Analysis of Variance
CACMS	Committee on Accreditation of Canadian Medical Schools
CanMEDS	Canadian Medical Educational Directives for Specialists
CaRMS	Canadian Residency Matching Service
CFPC	College of Family Physicians of Canada
CME	Continuing Medical Education
CQI	Continuous Quality Improvement
EFPO	Educating Future Physicians of Ontario
ICAR	Interprofessional Collaboration Assessment Rubric
ICEHR	Interdisciplinary Committee on Ethics in Human Research
ICU	Intensive Care Unit
IPE	Interprofessional Education
ITER	In-training Evaluative Report
LCME	Liaison Committee on Medical Education
LPN	Licensed Practical Nurse
MCC	Medical College of Canada
MD	Medical Doctor(ate)
MSF	Multi-Source Feedback
NL	Newfoundland and Labrador
OSCE	Objective Structured Clinical Examination
RCPSC	Royal College of Physicians and Surgeons of Canada
RN	Registered Nurse
WHO	World Health Organization

LIST OF APPENDICES

Chapter 1 Introduction

1.1 Medical Education in Canada

Canada's current medical education programming is guided by multiple national and international organizations. The standard pathway for an individual undertaking a career in medicine is directed as follows:

An undergraduate medical degree (M.D.), or international equivalent, must be completed at one of Canada's seventeen medical schools or other accredited international institution. In Canada, final M.D. examinations and distinctions are conferred by individual medical schools.

To practice as a physician in North America, regardless of where undergraduate medical education was completed, a licensing examination – administered by the Committee on Accreditation of Canadian Medical Schools (CACMS) in conjunction with the American-based Liaison Committee on Medical Education (LCME) – must be successfully completed.

Next, M.D. graduates must complete annual, progressive requirements within a residency program, based out of the same seventeen medical schools, which range in length from two years (Family Medicine) to six years (Cardiac or Neurosurgery). International medical graduates must complete an additional requirement to apply for a residency program in the Medical Council of Canada Evaluating Examination.

Finally, to become a licensed, practicing physician or surgeon, a resident must successfully complete a two-part, standardized, nationwide written and practical exam administered by the Medical Council of Canada (MCC). Residency certification examinations are administered by the Royal College of Physicians and Surgeons of Canada (RCPSC, or Royal College) for all specialties, excluding family medicine which is directed by the College of Family Physicians of Canada (CFPC). Medical licenses are then granted and regulated from individual provincial medical colleges.

1.2 The Royal College and CanMEDS

The Royal College, established in 1929, is a private, not-for-profit, national organization which sets standards for all medical institutions in Canada which certifies specialist distinctions – fellowships – for both physicians and surgeons, excluding family medicine. The Royal College's mandate is to *"strengthen specialty medicine to meet society's needs"* (RCSPC, 2013). It attempts to uphold this statement through development, administration, and supervision of accreditation procedures of medical institutions, examinations to certify specialists, maintenance of certification, and educational objectives (RCSPC, 2013).

Currently, medical education in Canada – undergraduate, post-graduate / fellowships and continuing medical education (CME) – is underpinned by an educational framework developed in 1993 and implemented in 1996 by the RCPSC entitled "*Canadian Medical Education Directives for Specialists"*, or more simply, CanMEDS. This framework describes seven core roles that specialists should demonstrate competency in: Medical

expert, Communicator, Collaborator, Manager, Health advocate, Scholar, and Professional (RCPSC, 2005). Visually, the CanMEDS roles, and their interconnectedness, are displayed in a RCPSC trademarked 'flower' or 'daisy' image. The image clearly depicts Medical Expert as the central role encircled by the remaining six roles acting as supporting competencies.

1.3 The History of CanMEDS

The creation and development of the CanMEDS roles and subsequent competencies arose from an interesting history. A thematic discourse analysis was completed by Whitehead, Austin, and Hodges (2011), using archival documents from the University of Toronto's Thomas Fisher Rare Book Library, describing key events that lead to the advent of the current CanMEDS system. The following is a synopsis of their findings.

In 1987, a project entitled *Educating Future Physicians of Ontario* (EFPO) commenced following the 1986 Ontario Physicians strike. The striking physicians were protesting the federal government's legislation to ban "over- or extra-billing" (i.e., a physician billing a patient for a service that could be billed to Medicare or medical insurance company). It is thought physicians were conducting such practice in silent protest of their disagreement with their salaries. Conversely, the federal government felt that the healthcare of Canada should belong to the citizens and not to physicians. As such, the public was losing faith in the values that physicians were upholding. The twenty-five day strike was poorly supported by both physicians and the public. Eventually, the legislation to ban over-billing was carried out.

In response to the obvious divide between Ontario physician and patient perspectives on healthcare in Canada, the EFPO began to investigate a method to improve physician's ability and preparedness to meet the societal need from an educational standpoint – starting at the undergraduate level. This process began with obtaining public input as to what roles *they* thought physicians should embody. Over years of extensive surveying, consultations, and iterations the initial EFPO roles were slowly transformed into the CanMEDS roles utilized today (Figure 1).

Figure 1 - Progression of Physician Roles from EFPO (1987) to CanMEDS (1996)

Distinct from the Royal College, the speciality of Family Medicine has its own professional organization, the College of Family Physicians of Canada (CFPC), which provides educational objectives, examinations and accreditation to institutions offering a residency program in Family Medicine. Closely linked to the CanMEDS roles are CFPC's Four Principles (CFPC, 2013):

1. The family physician is a skilled clinician
2. Family medicine is a community based discipline
3. The family physician is a resource to a defined practice population
4. The patient-physician relationship is central to the role of the family physician.

1.4 CanMEDS Collaborator Role

The Royal College defines physician collaboration as "*effectively working within a health care team to achieve optimal patient care"* (RCPSC, 2006). The two 'key competencies' for this role require that physicians are able to:

1) *Participate effectively and appropriately in an interprofessional health care team*.

2) *Effectively work with other health professionals to prevent, negotiate and resolve interprofessional conflict*.

Recently, in 2012, the Royal College published another handbook offering recommendations specifically on the CanMEDS Collaborator Role: "*The CanMEDS Toolkit for Teaching and Assessing the Collaborator Role".* Although the majority of the book addresses developing the teaching of collaboration, it also provides insight into

selecting the appropriate collaborative assessment and evaluation tools. Two possible Collaborator assessment instruments are included. First, a general in-training evaluation report (ITER) focused on specific collaboration competencies such as 'participating in health care teams' or 'managing conflicts and differences' and secondly, "encounter cards" for specific interactions such as discharge planning or family meetings (RCPSC, 2012).

Assessment tools such as these often consist of a Likert scale of varying numbers of points and types of categories used to assign a score to a specific evaluative statement, or item. For example, an assessment tool utilizing a Likert scale may contain a five-point scale where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree (Likert, 1932). A thorough discussion on scales is found in the literature review.

The Royal College *CanMEDS Assessment Tools Handbook* offers further suggestions for possible assessment tools on the Collaborator role, including: written tests, reflective journals, video recording playback, multi-source feedback, and objective standardized clinical examinations.

1.5 Interprofessionalism in Healthcare

Prior to late-20th century, the healthcare industry was considered a *paternal* state, where physicians were the primary decision makers for treatments their patients would receive. Today's view is one of patient autonomy where physicians provide their patient adequate information regarding their medical issue, treatment options, and provide informed consent for future treatment. Further distancing the healthcare industry from paternalism

and coinciding with increased patient autonomy was the advent of interprofessionalism. The health care that patients receive today may include direct input from some, or all, of – for this research's purposes – three professional categories: *physicians* (including undergraduate and graduate medical students as well as practicing physicians), *nurses* (including registered nurses – RNs, licensed practical nurses – LPNs, and nurse practitioners), and *allied health professionals* (including, but not limited to, social workers, physical therapists, occupational therapists, technicians, pharmacists, speech language pathologists, dieticians, and pastoral care).

Successful patient care, the obvious mandate of any healthcare system, must incorporate satisfactory input from a multi-disciplinary team. Any lone member of the health care team cannot provide the tools necessary for appropriate patient care. To improve and uphold quality care for the patient and their family, as well as to promote long-term development of interprofessionalism, necessary educational structures must be initiated and maintained at all levels of medical education. Overeem et al (2009) reinforce that barriers to medical resident improvement in collaboration occur primarily due to the failure of hospitals to create a climate that is conducive to collegial support and lifelong reflective learning.

The delivery of interprofessional education (IPE) is a necessity at all phases of a healthcare worker's career. At Memorial University's Centre for Collaborative Health Professional Education, for example, IPE is offered for all undergraduate healthcare students as well as for individuals already working in their professional field. Students in

medicine, nursing, pharmacy, and social work meet for eight half day sessions and four half day modules to discuss various case studies involving interdisciplinary roles and perspectives. Surveys are distributed at the beginning and end of the year to investigate changes in the students' perceptions in the importance of interprofessionalism. At the professional level, collaborator care exists but in an inconsistent manner. Interprofessional health care staff, as well as medical students and residents, on some medical / surgical units in Eastern Health (a healthcare authority in Newfoundland) meet at regular intervals (either daily or weekly) to collaboratively discuss healthcare plans for patients under their care during that time frame.

1.6 Assessment in Post-Graduate Medical Education

Adoption and integration of the CanMEDS educational framework into post-graduate medical institutions demands the incorporation of appropriate assessment techniques and evaluative practices. For each of the seven roles, the Royal College provides recommendations for assessment methods in their 2006 publication "*CanMEDS Assessment Tools Handbook"*. The Royal College recommends that ideal evaluation should involve a multifaceted approach at varying time intervals to assess different aspects of skill, attitude, behavior, and performance (RCPSC, 2006). In support, Massagli and Carline (2007) note that physician competence is multi-dimensional and that no single tool will be able to assess all aspects of competence, yet many institutions attempt to do so through use of integrated ITERs where all CanMEDS role are assessed together.

Generally, CanMEDS roles are to be assessed in both clinical and non-clinical settings through various techniques suggested by the Royal College. Clinically, the RCPSC recommend methods including ITERs, observed procedures, 360-degree feedback, or chart audits. Non-clinical methods include written tests, objective standardized clinical examinations (OSCE), rotations focused on a specific role (such as Manager or Scholar), logbooks, portfolios, or observed teaching. Some assessment techniques suit some roles better than others. Methods used will vary across institution as each medical faculty will have different needs and expectations of their resident's medical education during various points of their residency.

The research completed in this thesis primarily focuses on assessment of the CanMEDS Collaborator role through an ITER-based 360-degree assessment.

1.7 Assessment vs. Evaluation

The ICAR is primarily intended to be utilized as a formative assessment tool to help identify and address gaps in a medical student's collaborative ability but it could also be used for evaluative purposes at the program director's discretion. The terms 'assessment' and 'evaluation' are often used synonymously outside of the education world. Although similar, the terms are describing two unique aspects of education. The Newfoundland and Labrador Government's Department of Education defines Assessment and Evaluation as follows (via a publication entitled, *Assessing and Evaluating Student Learning*):

Assessment: "*the process of gathering information on student learning"*.

Evaluation: "*the process of analysing, reflecting upon, and summarising assessment information, and making judgements and/or decisions based on the information collected"*.

Assessments may be classified as either 'formative' or 'summative'. Formative assessments are used as educational tools as a learner progresses throughout an educational pathway. A formative assessment is conducted with the intention of both educator and learner utilizing the results or feedback to identify and address gaps in their knowledge, skills, and competencies (Dent and Harden, 2005). Formative assessments are often provided without a quantitative measure being supplied for quality of work. Summative assessments are similar to formative assessments in format but focus on providing a measure of comprehension in a learner with a desired set of competencies to be achieved (Dent and Harden, 2005).

1.8 Issues with Assessing the CanMEDS Roles

The primary objective of this research project aims to fill a void in the evaluation process of CanMEDS roles in post-graduate medical education. Despite the publications (i.e., *CanMEDS Assessment Tools Handbook*) and the recommendation (i.e., repeated, multi-faceted, and multi-tool approaches) there remains a gap in providing appropriate assessment methods for the six non-'Medical Expert' CanMEDS roles. Logically, the Medical Expert role is the most objective of all the roles to assess and as such is the most assessed role across Canadian universities (Chou et al, 2008). At the bottom of the list, provided by Chou et al (2008), of least assessed roles, and with the one of the lowest

satisfaction scores, was the Collaborator role. The Collaborator role was assessed 2.5 times less frequently than Medical Expert and also received lower satisfaction ratings by program directors in assessment quality than the Medical Expert role. The same trends existed for Health Advocate, Manager, and Professional.

For example, at Memorial University in Newfoundland, Canada, all non-medical expert CanMEDS roles are evaluated through an ITER using a 5-point Likert scale (1 = rarely meets to 5 = consistently exceeds reasonable expectations) based on the core competency statements listed by the RCPSC (Appendix A). More specifically, the Collaborator role is assessed on only *two* competency statements, which are the key competencies directly quoted from the CanMEDS framework:

1) T*he ability to participate effectively and appropriately in an interprofessional healthcare team*

2) *The ability of effectively work with health professionals to prevent, negotiate, and resolve interprofessional conflict*

As Massagli and Carline (2007) noted, evaluating a learner's ability for a required skill using a limited number of assessment points underscores the need for improvement in this process. Our research intends to eventually increase the depth of assessment for the Collaborator role and further the growth in interprofessional education and teamwork at Canadian medical education faculties.

1.9 Multi-Source Feedback

An ideal method to obtain an accurate evaluation of an individual's performance - in any environment - regarding any skill, attitude, knowledge, behavior, etc., is to accumulate and analyze the highest quality data possible. One such method would be to combine assessments from multiple perspectives. This survey / questionnaire-based evaluation method is commonly named 'multi-source feedback (MSF)' or '360-degree assessment or evaluation'. As health care is unquestionably an interprofessional industry, MSF is becoming increasingly integrated into the healthcare industry for assessment of both medical learners and faculty (Overeem *et al*, 2009). As such, MSF is seen as an effective method leading to positive impacts from feedback delivered to residents (Stark *et al,* 2008).

The origins of MSF date back over half a century. Fleenor & Prince (1997) describe how MSF evolved from initial development by militaries during World War II, to organizations gathering employee opinions on various internal issues via surveys, to individual assessment in the 1980s, combating the traditional single source, top-down assessment. The healthcare industry adopted MSF methodology in the late 1990`s as a method of improving the care provided by its employees at all levels (Lockyer, 2003; Overeem *et al*, 2009). The growth and utilization of MSF in the business world is obvious as companies such as 3Dgroup (www.3dgroup.com) and STAR 360 Feedback (www.star360feedback.com) exist to conduct 360-degree assessments of an organization's employees. As figure 2 depicts, MSF is utilized in the business realm

primarily as a development tool for employees and management. Ensuring that personnel meet desired standards accounts for 20%, while pay and promotion is attributed for the final 10% of usage.

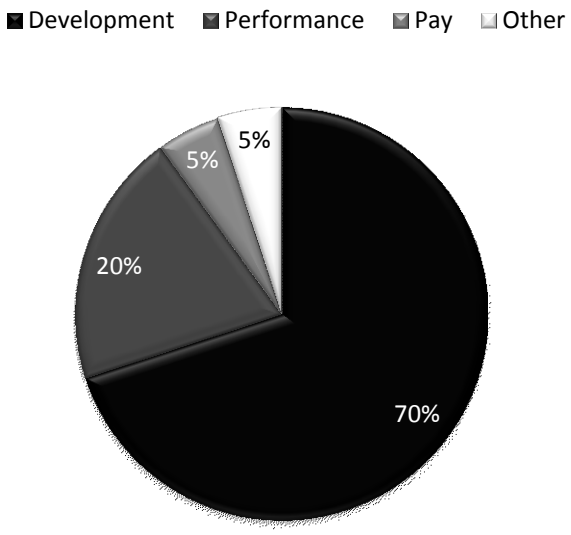


Figure 2 - Utilization of MSF Procedures in Businesses (Hewitt Associates & Nowack, 2011)

One example of MSF in practice in a health care setting may be as follows: an undergraduate medical learner could be assessed using an appropriate ITER during, or after, a particular rotation by, any or all of, their attending physician, junior and/or senior resident, patient care co-ordinator, nurses, allied health professionals, patients and families, and even themselves. A program co-ordinator could then collect all assessments to analyze the strengths and weaknesses of the learner. Feedback could then be provided to the learner in a face-to-face discussion or through a summary report.

1.10 Education Meets Clinical Epidemiology

There is some initial skepticism when medical education research is presented under the umbrella of clinical epidemiology. Some confusion is valid considering Parfrey and Barrett (2009) define clinical epidemiology as "the science of human disease investigation with a focus on diagnosis, prognosis and treatment". Realistically, the goal of clinical epidemiology, much like medical education, is to improve health care for the individual patient and general population.

Through various research designs, from retrospective case-control studies to prospective cohort studies, clinical epidemiologists attempt to determine which *exposure* (i.e., smoking) can lead to a specific *outcome* (i.e., lung cancer). From this, statistical analyses can be applied to calculate a multitude of confirmatory and prognosticating statistics. These statistics suggest relationships, likelihoods, effect sizes, and significant differences between samples of specific populations.

Authors Parfrey and Barrett have pioneered the translation of clinical epidemiology methodology to the genetics world. From a medical education standpoint, these same epidemiological techniques can be utilized to determine the likelihood or basic success of a specific educational intervention (*exposure*) on a medical learner's knowledge, skill, or behavior (*outcome)*. Countless research studies attempt to demonstrate a resident's ability to retain knowledge, skills, attitudes, or behaviors after a specific intervention is utilized. For example, clinically, will the use of high-fidelity simulation improve a resident's skill at surgery better than a low-fidelity simulation (Tan *et al*, 2012)? A non-clinical example

could include investigating whether the use of journal entries in residency yield a more self-reflective physician than no journal writing (Webb and Merkley (2012).

In the case of this study, the main purpose is to determine the reliability of a new, thorough, innovative ITER to be used in the future as an *exposure* to produce a future *outcome* of becoming a better collaborator and interprofessional team member.

1.11 Research Goals

1. Determine inter-rater reliability of Interprofessional Collaborator Assessment Rubric (ICAR) and compare ICAR scores between rater groups through multi-source feedback.
2. Investigate the feasibility of incorporating the ICAR in medical resident assessment. E.g., Will the ICAR be a tool that health professionals use to rate residents.
3. Investigate the resident's perceptions of the ICAR and MSF.
4. Determine if evaluation biases (gender, years of experience, etc.) exist when assessing collaboration in residents.

Chapter 2 Literature Review

2.1 CanMEDS

In Canada, as previously mentioned, the CanMEDS framework began development in 1993 and was implemented in 1996 by the RCPSC outlining the seven key roles for specialist physicians. In recent decades, there has been a paradigm shift in graduate medical education. A previous focus on process and structure has transformed to one on product, or outcome, based approaches (Musick *et al*, 2003). This is evident from the creation and development of physician core competencies that national accreditation councils recommend. The CanMEDS principles were tangibly successful in this manner as several countries including Denmark, The Netherlands, Australia and New Zealand have adopted the CanMEDS framework as the focal point of their own medical education (Ringsted, Hansen, & Scherpbier, 2006). The United States adopted a similar approach shortly after CanMEDS as the Accreditation Council of Graduate Medical Education (ACGME) initiated their Outcome Project in 2002 (ACGME, 2012). The Outcome Project identifies six general competencies corresponding with the CanMEDS roles to improve resident education and assessment: patient care, medical knowledge, professionalism, systems-based practice, practice-based learning and improvement, and interpersonal and communication skills.

2.1.1 Role Inequality

Although each CanMEDS role contributes vital importance to providing adequate healthcare, are all roles equally important? If not, which role(s) are of the *most*

importance? More specifically for this thesis' research, where does the Collaborator role rank in importance? The manner in which medical schools and residency programs answer this question could have tangible effects on health care.

Four studies in particular attempted to deduce this question. First, Stutsky et al (2012) distributed surveys to all practicing physicians in four provinces (British Columbia, Alberta, Saskatchewan, and Manitoba) and the three territories. The survey consisted of 21 questions on a 5-point Likert-style scale ranging from a negative response to a positive response. The same three questions were asked for each CanMEDS role measuring the physician's attitudes toward each role's complexity, frequency, and criticality. Eighty-eight surveys were completed yielding a response rate of only 3%. The results showed that the highest priority role was Medical Expert, followed by, in order, Communicator, Professional, Collaborator, Scholar, Manager and Health advocate.

A second study in Denmark by Ringsted et al (2006) also used a 21-question survey (three statements per role) measuring physicians attitudes towards the importance of and their confidence in a specific CanMEDS roles and competencies using a 5-point Likert-style scale where 1 = totally unimportant and 5 = very important. The 42.8% response rate had 3476 physicians complete the survey. The results were subdivided by responder title: intern (one year post-MD), introductory year resident trainee, resident, and specialist. Results showed no difference in the importance between roles across the responder groups. This study's findings listed the Communicator role as the most important, while Collaborator and Health Advocate roles were noted to be of the least

importance. With regards to confidence in specific roles, on average, Communicator scored the highest followed by Scholar, Collaborator, Professional, Health Advocate, Medical Expert, and Manager.

A third study by Chou et al (2008) conducted a web-based survey (5-point Likert scale) of Canadian residency program directors to determine the assessment tools, and thus the level of assessment, of each of the CanMEDS roles at their institution. The 53.2% response rate from 280 directors yielded two important findings. First, with respect to the total number of assessments provided per role, Medical Expert received the highest ranking followed by Communicator, Scholar, Professional, Health Advocate, Collaborator and Manager. Secondly, the residency program directors provided their level of satisfaction with their program's evaluation of each specific CanMEDS role. Medical expert ranked first, followed by Communicator, Scholar, Professional, Collaborator, Manager, and Health Advocate.

Finally, Arora et al (2009) similarly asked surgery residents and attending surgeons in London, UK to score the importance of each of the CanMEDS roles. The 74% response rate yielded 92 responses indicating significant findings and potential educational outcomes. The Collaborator role scored poorly in importance for all physicians. First year residents rated Collaborator second lowest in importance after Manager; junior residents rated it third lowest, and senior residents rated it lowest importance and, finally, attending surgeons also ranked it second lowest. The study also asked the same participants about their perceived competence in each role. Surprisingly, none of the 92 physicians felt they

had ‘mastered’ the Collaborator role. All four of these studies demonstrate a low level of importance, competence, and assessment regarding the Collaborator role.

Thus, it may be logical to assume that if collaboration is poorly addressed from a medical education perspective then interprofessionalism – collaboration in practice settings – may be struggling as well. Specific questions need to be addressed, such as: how physician-centric are medical wards? Do nurses, allied health professionals, patients or family have input in the assessment of medical learners? Specifically to this study, to what extent would physicians, nurses, and medical professionals agree if, and when, they were to all evaluate the same medical learner?

2.2 In-Training Evaluation Report (ITER) Development

To ensure competent physicians and surgeons graduate from residency programs, in-training evaluation is an educational initiative utilized to measure a resident’s ability in a particular skill set throughout, or upon completion of, a specific rotation or residency year. An in-training evaluation report, or ITER, is the most commonly used form of in-training evaluation. Chou *et al* (2008) found that 92% of medical education programs use ITERs followed by oral examinations (86%) and multiple-choice questionnaires (MCQ) (72%). ITERs usually appear in the form of a multi-category Likert scale, global rating scale, or objective dichotomous checklist; among other possible formats (Gray, 1996). ITERs are largely (86% of 149 residency programs) constructed by an individual program and often (58%) custom designed for individual rotations (Chou *et al*, 2009). ITERs are utilized at varying time intervals during a residency. Chou et al (2009) found that

Canadian medical faculty's used ITERs monthly (32%) , bimonthly (27%), and quarterly (40%).

Despite the ubiquity of ITERs, the literature has identified several issues with their implementation in medical education. The most critical points on a rather exhaustive list of issues to be further discussed include: appropriate assessment tool construction, lack of standardization, lack of rater training, limited reproducibility multiple errors in rating scales, subjectivity, limited observation of medical learner performance, and inaccurate recording of observations, and a lack of multiple perspectives. Fundamentally, this array of issues can be categorized into three general problem areas: the assessment tool, the assessment process, and the assessor(s).

2.2.1 The assessment tool

Before any evaluation process can begin, the necessary assessment tools must be constructed. The construction of an ITER is an arduous process entailing many steps from determining the number and content of evaluative items to the size of the scale – number of categories – to acquiring both validity and reliability of the tool before implementation as an acceptable evaluative tool. Any deviation from the previously listed steps could lead to invalid assessments which may potentially impact a learner's education and career as well as the residency program.

Validity, in general, is defined as the extent to which an item or instrument measures what it intends to measure (Parfrey and Barrett, 2009). Initially, an instrument must possess both *face* and *content* validity. Face validity measures whether the items, or

evaluative statements, appear to measure what the tool is intended to measure. This is often confirmed through content validity in which evaluators or experts can suggest specific items to be removed from the tool if deemed irrelevant, confusing, or likely difficult to observe. Next, the remaining items must possess *construct* validity, which is observed when a predictable result can be produced. For example, a senior resident should score higher than a junior resident on an ITER measuring skills for more advanced techniques. Finally, if possible, *criterion* validity can be possible if the instrument can be compared to a current 'gold standard' measuring the same attribute. Criterion validity is more difficult to prove when developing new assessment tools. In such a case *concurrent* validity may be useful, as you may compare the scores produced from the new assessment tool to scores of an existing tool (McMillan and Schumacher, 2006).

Closely associated with, but mutually exclusive from, validity is reliability. Reliability, in general, is defined as the ability to produce the same or similar results during measurement over different occasions or using different observers (Streiner and Norman, 1989). Reliability mainly presents itself in two forms: *test-retest* reliability and *internal consistency* reliability. The former occurs while comparing scores of a learner assessed multiple times, using the same instrument (McMillan and Schumacher, 2006). The latter is a correlation between the different items within the instrument as each item should be probing a different aspect of a common attribute (i.e. collaboration skills). Reliability is measured via coefficients ranging in value from 0 to 1, where 0 equals no correlation or reliability and where 1 equals perfect correlation or reliability. For internal consistency reliability, Cronbach's alpha (α) is the accepted statistic with a value of >0.7 indicating

suitability (McMillan and Schumacher, 2006). There are many different coefficients used for test-retest reliability ranging from Cohen's Kappa, Fleiss' Kappa, Intra/Inter-class correlations, and generalizability coefficients (to be discussed further in section 2.7).

A second problematic area in effective assessment tool implementation is dealing with errors in rating scales during the development process. A researcher may believe that any scale may work with their new questionnaire idea. But, how do they know if they should choose a 4-, 5-, 7-, 9-, or 10-point scale? Will they use an expectation, frequency, agreement or visual analog scale? To achieve the desired result the proper scale must be used. To specifically address this issue in the medical education world, Streiner and Norman (1989) authored a book entitled *Health Measurement Scales: A Practical Guide to Their Development and Use*. The authors suggest that the minimum number of categories to be used should be five to seven. With this, they address 'end-aversion bias' or 'central tendency bias' where raters are less inclined to respond to the extreme values (1 or 7). They rationalize that people have difficulty in making absolute judgements and tend to feel more comfortable staying close to the middle value. This indirectly creates a smaller scale, and loss of sensitivity and reliability. To avoid this issue they suggest increasing the size the scale by two points; one on each end. For instance, if the researcher wishes to use a 5-point scale, they should consider using a 7-point scale instead. As well, they recommend avoiding absolute statements, such as 'never' or 'always', in the item or question itself. Furthermore, a researcher must consider whether to employ an even or odd numbered scale. Again, the choice will depend on the

researcher's desired outcome as an odd number of categories allow a 'neutral' response while an even number forces a positive or negative response.

Next, the type of scale that is used depends on the type of response that the researcher desires. Each scale type has its own advantages and disadvantages. Assume, for this thesis' purposes, they desire a continuous, not a dichotomous or free text, response. A major issue with choosing a proper type of scale is how the categories will be interpreted by the rater. For instance using a frequency scale, rater A may define the term 'rarely' as almost never occurring while rater B may define it as merely infrequent. Thus, regardless of the scale that is used, it is essential to improve validity and reliability by providing definitions of the category terminology. Expectation scales are likely to be interpreted differently by each rater as one's expectation may differ from the next. The visual analog scale – usually a 100 mm line with two extreme categories only - attempts to eliminate such limitations as the raters are not limited by multiple categories and definitions but it often leads to lower reliability values due to an infinite amount of categories.

A final point on assessment instrument construction centers on the eventual statistical analysis to be completed upon data collection and entry. It is imperative that the choice of analysis used is based on the level of data collected, otherwise the ensuing results have an increased risk of yielding a false positive or false negative result (Jamieson, 2004). For instance, if using a yes/no scale then statistical analysis is limited to specific tests such as logistic regression. However, if using a common 5-point Likert Scale, ranging from 'strongly disagree' to 'strongly agree', then concern arises regarding how to analyze the

data. Although the data will appear as continuous data, the numbers themselves are merely representing categories; 1 = strongly disagree, 3 = neutral, 5 = strongly agree. Thus, the ensuing dilemma is whether to consider the assessment data as ordinal (intervals between numbers are unequal) or as interval / ratio (intervals between numbers *are* equal). This debate is heavily documented in the literature with proponents backing both classifications (Jamieson, 2004). Those supporting the designation of ordinal data, and subsequent appropriate analyses, suggest that the rater's perception of the difference between categories *agree* and *somewhat agree* cannot be assumed to be equal to the difference between *agree* and *strongly agree* (Jamieson, 2004). However, this *is* an assumption of interval data. For instance, regarding a Celsius temperature scale, the difference between $10^{o} – 20^{o}$ and $20^{o} – 30^{o}$ is presumed to be equal. However, according to Blaikie (2003) and Cohen (2000) researchers often allow this ordinal data based assumption to persist. Those who statistically analyze Likert-type scale as interval data propound that a sufficiently large sample size and normal distribution should trump level of data (Jamieson, 2004). The divide between the two statistical views will undoubtedly persist, but the ethical and responsible researcher will hopefully choose the correct analysis which coincides with the level and quality of data.

2.2.2 The assessment process

Once an appropriate assessment tool has been constructed, a researcher must contend with issues arising during the assessment period. A glaring obstacle is developing a method for training prospective raters with practice, knowledge, or familiarity with the

instrument. As mentioned above, individuals will have personal definitions regarding terms used in scale categories. For example, there is ambiguity in terms such as 'slightly', 'somewhat', 'strongly', 'rarely', 'often', etc. Chou et al (2009) found that only 25%, of the 149 participating residency programs in their study, offered ITER training before assessments were completed. This allows for personal habits, views, or biases to affect assessor ratings, which may reduce the reliability and validity of the instrument. Providing a form of training, or at least specific definitions of terminology used in the tool, should minimize such bias.

There are also potential obstacles from an administration perspective. Researchers must decide whether to distribute the assessment tool as a paper document or as an online document which raters may submit electronically. In either format, participants may complete the entire survey, or choose to partially complete the assessment or not respond at all. A 2003 study by Sax et al. found that web-based tools provided lower response rates than paper-based tools (24.0% vs. 17.1%, no *p*-value provided).

2.2.3 The assessor(s)

Finally, a large obstacle in achieving the most valid results upon distributing an assessment tool involves the assessors or raters. Specifically, using the example of resident assessment for this discussion, the relationship between the assessor and resident may influence results. Streiner and Norman (1989) discuss the following rater biases that should be accounted for during the assessment process. The 'halo effect', first described by Johnson and Cujec (1989), occurs when the resident's rating is influenced by the

rater's perception of the resident. For example, a resident's low rating on sub-par clinical skills lead to a low rating on their non-clinical skills. Another bias, 'positive skew,' exists as raters are often uncomfortable with providing negative evaluations of individuals, thus forcing an unbalanced result toward positive responses. The resulting 'ceiling effect' may cause difficulty in observing a resident's improvement over time. Similarly, 'yea-saying' or 'nay-saying' bias exists when raters provide all positive or all negative responses, respectively, regardless of the statement. Assuming that the raters ignore the content of the item, all of the above can be identified by 'reverse-wording' specific statements. Structurally, these 'reverse-worded' statements include a negative term transposing the scale where a low score, 1, would now indicate a favorable score and a high score, 5, would now indicate a less favorable score. Finally, the tendency for raters to avoid extreme ends of the scale is called 'end-aversion bias' or 'central tendency bias', which essentially reduces the range of possible scores.

2.3 Interprofessional Collaboration and Education

Moving from the creation and development of satisfactory assessment tools, we turn to the specific focus of this study: providing an appropriate, reliable, and valid assessment tool to aid in evaluating a resident's collaboration ability.

As the age of paternal, physician-centric healthcare began to fade in the mid-to-late twentieth century, the movement toward a team of multiple healthcare professionals cohesively working to improve patient care began to grow. As healthcare and medication improves leading to increases life expectancy, as well as need for complex care, there is a

larger burden placed on the healthcare system and the individuals who work within it. A 2006 World Health Organization (WHO) study further exacerbates this issue as they found major worldwide deficits in healthcare personnel. Specifically, they estimated that the workforce in the 'Americas' needs to increase by 40% just to reach a minimal standard (WHO, 2006). The rising burden and expectation placed on all healthcare professionals, in addition to the lack of personnel, requires a greater emphasis on functional interprofessional collaboration. However, high quality collaboration in high-stress environments may not come naturally. To underscore this point, in 2002 Health Canada published a report entitled *Building on Values: The Future of Health Care in Canada* which was commissioned to investigate the current status of Canada's health care system and provide insight and recommendations for its future. The 360-page report outlined many concerns with the Canadian healthcare system but of specific interest was the need for new approaches addressing collaboration and education among health care professionals to maximize efficiency. One clear example of the benefits of enhanced interprofessional collaboration is in the field of patient safety. Manser (2009) conducted a literature review investigating the effect teamwork in "highly dynamic domains of healthcare" (i.e., the emergency room and intensive care units) has on patient safety and quality of care. Manser reviewed one hundred and one studies between 1998 and 2007 and came to the conclusion that poor communication and teamwork were among the leading contributing factors (22-36% of reports) of adverse events.

The WHO defines interprofessional education (IPE) as the learning resulting from the interaction between two or more professionals with the intent to improve health outcomes

at a multitude of levels (WHO, 2010). As previously mentioned, IPE can, and should, be addressed at all levels of a healthcare professional's career. The WHO's 2010 report indicates the most successful approach to improving interprofessional collaboration is by developing an integrated educational and clinical culture that "commits to" and "champions" interprofessional education. Some measures they indicate that will aid in creating such a culture are staff training, institutional and managerial support and commitment, compulsory attendance, contextual learning, and appropriate assessment. Our research aims to supplement the area of assessment in the continuing development of interprofessional collaboration culture.

2.4 Interprofessional Collaborator Assessment Rubric (ICAR)

Interprofessional learning environments currently recognize the need for competency-based evaluations of medical learners (Davis & Harden, 2003). Beyond a medical learner's clinical skills, it is essential to evaluate interpersonal, interprofessional, and problem solving skills as required traits to create and promote excellence in health care (Verma *et al*, 2006). Similarly, to demonstrate the aforementioned skill, residents must be able to effectively exchange information with colleagues, patients, family, and professionals from all other medical disciplines (Joshi, 2004). Consequently, the need for a valid and reliable interprofessional assessment for these competencies is essential in the medical education – and professional – environment.

Developed by Dr. Vernon Curran *et al* in 2011, the Interprofessional Collaborator Assessment Rubric (ICAR) intends to fulfill the growing demand for either formative or

summative assessment for CanMEDS Collaborator role. The original version of the ICAR (Appendix B) contained 31 evaluative items divided into 6 domains. The evaluative items are based on the competencies of the collaborator role set out by the RCPSC CanMEDS framework. Each category statement is evaluated on a scale of 1 to 4 (1 = minimal, 2 = Developing, 3 = Competent, 4 = Mastery) based on the frequency of demonstrated ability of the medical learner outlined by behavioral indicators. For example, a learner scoring in the *Developing* category would demonstrate the desired trait *occasionally.* The rubric also contains a 'Not Observable' column for the assessor if the interaction doesn't allow the resident to display that behavior. A comment section is available for additional notes on the learning encounter or if the assessor has trouble evaluating any competency statement.

The ICAR was developed through a two-stage mixed methods approach. The first phase involved the validation of a set of collaborator competencies relevant to various interprofessional learning environments and the CanMEDS framework. An extensive literature review, including the grey literature, was conducted to analyze and determine the most effective competencies, descriptors, and statements for the ICAR (Curran *et al,* 2011). The second phase included obtaining expert opinion via a two-round Delphi survey. A pan-Canadian group of English and French speaking experts in interprofessional education and collaboration were asked to assess the ICAR for validity. Multi-site focus groups, involving faculty and students from health care professions, were utilized to evaluate clarity and practicality of the instrument (Curran *et al*, 2011). The ICAR requires reliability testing – the primary outcome of this thesis – in the professional field for future use across medical faculties.

2.5 Multi-Source Feedback (MSF)

If a residency program deems a resident to be unsatisfactory in a specific competency, then the necessary pathway must be followed to allow a positive, supportive change in the resident's performance and behavior. This concept of continuous quality improvement (CQI) is paramount in medical faculties and national associations such as the RCPSC and CFPC. Through CQI healthcare management, staff, and professionals all work towards a common goal of exceeding the expectation of the patient and their family. Overeem et al (2007) conducted a systematic review of daily performance assessment in eight Dutch hospitals finding that two-thirds of physicians believe a change in behavior will occur when quality, valued feedback is appropriately provided and that negative, poor, or inconsistent feedback does not spur the necessary positive change. The researchers underscored that assessment-driven learning is increasingly acknowledged as a necessary principle of medical education. As such, to increase feedback quality, a residency director, for example, must be provided with adequate assessment data to effectively evaluate a resident's performance and determine areas of strength and weakness.

Since 1999, to ensure comprehensive assessments were being conducted on learners, medical faculties began to borrow 360-degree assessment or multi-source feedback (MSF) techniques from the business and industrial world (Overeem et al, 2009; Ogunyemi, 2009). As previously defined, MSF is an assessment technique which incorporates the perspective of multiple sources of observation regarding a learner. As of 2009, a decade later, there were over 4000 residency programs in North America and

United Kingdom using MSF to assess residents and fellows (Overeem et al, 2009). MSF feasibility, reliability, and validity has been studied in various medical speciality programs including, but not limited to: Emergency Medicine (Garra et al, 2011), Internal Medicine (Warm et al, 2010), Obstetrics / Gynecology (Joshi et al, 2004), Pathology (Lockyer et al, 2009), and Psychiatry (Violato et al, 2008).

Lockyer and colleagues from University of Calgary in Alberta, Canada have been instrumental in researching, developing and promoting successful MSF practices in healthcare environments. In a 2003 study they reported that effective MSF appears to be dependent on a multitude of departmental factors including organizational support, creation of steering committee, continual monitoring, and valid and reliable instrument design and testing. Furthermore, they suggest participants, both ratees and raters, must understand the purpose and goals of the MSF process and how it will be of value to the individuals, the residency program, and the healthcare in general. Massagli and Carline (2007) and Campbell *et al* (2011) note that MSF is best utilized when incorporated as a formative process whereby residents can review the results, or are provided feedback, to develop a plan of action to reach competency with their mentor or residency director. As well, raters tend to provide more accurate and less lenient ratings when MSF is used as formative, rather than summative, evaluation.

A selling feature of MSF is in its flexibility as any aspect of an individual's performance may be assessed assuming the appropriate valid instrument and assessors are recruited. Here, in this thesis, the focus is on applying MSF as a preferred tool for evaluating

resident's performance in more subjective – colloquially referred to as 'soft' – competencies, specifically the Collaborator role. These subjective or soft competencies commonly include humanistic qualities such as the CanMEDS roles of Collaboration, Communication, and Professionalism. The belief persists that it is only through strong interpersonal skills that a physician can become a truly effectively part of the healthcare team (Joshi, 2004). Tertiary care hospitals utilize a high degree of interprofessional engagement to provide successful patient care. Thus, MSF may be an adequate assessment tool in helping residents improve upon deficiencies through feedback from other members of the healthcare team, as well as patients and their families (Massagli and Carline, 2007)

2.5.1 Advantages of Multi-Source Feedback

A surge in popularity and implementation of MSF in medical faculties is attributed to the numerous advantages over single-source assessment and feedback. Numerous medical faculties across North America and Europe have studied the effectiveness, feasibility, and positive impact of 360-degree evaluations. Hammock *et al* (2007) suggest MSF has been one of the most important mechanisms that influence the delivery of interprofessional education by increasing awareness of the roles of other medical professionals that contribute to quality patient care. Wood *et al* (2006) conducted a review of the literature on MSF regarding its application in healthcare and concluded, among other points, that incorporation of multiple perspectives in various environments is essential to evaluate performance. Participating residents felt that the evaluations increased their awareness of

how they interacted with patients (Wood et al, 2004). When ratees take part in the evaluation process, it allows self-reflection, increased engagement in the evaluation process, and comparison as to how their self-assessment aligns with the perceptions of those they interact with. Similarly, it allows assessment from perspectives which may rarely offer input, such as nurses, allied health professionals or even patients – to be discussed further.

From a feasibility perspective, Joshi et al (2004) demonstrated that in a stable institution, with a relatively small number of residents, MSF is a practical, effective evaluation of interpersonal and communication skills. Furthermore, Overeem et al (2007) investigated the amount of time spent per assessment method. They found that MSF consumed an average time of one hour of the assessor's time in comparison to other more time-intensive methods such as portfolios – a collection of a resident's work in a specific CanMEDS role used to set goals and track progression – which may take up to 15 hours per assessor.

2.5.2 Disadvantages of Multi-Source Feedback

Despite many advantages, MSF should be not be viewed as a panacea of assessment woes as the process has its own wealth of challenges to overcome. At the 2011, International Personnel Assessment Conference, psychologist Dr. Kenneth Nowack presented on MSF as being a less-than-suitable assessment method. Much of his argument, as follows, is supported by the medical education literature.

First, as evaluation is a necessity for all medical students (undergraduate and graduate), there is a constant demand on attending physicians to participate in assessment. However, increasing the burden of evaluators either by sheer number of unique assessments or the length of an individual assessment may lead to survey fatigue (Porter et al, 2004). This has the effect of lowering response rates and limiting potentially useful feedback for the learner. Although for each assessment the overall time-investment from the rater may be low, there is a larger burden applied to the administration of the evaluation. To ensure high response rates, hand-delivered paper surveys are often distributed compared to the less-responded-to online survey. However, effective MSF may be hindered by data collection procedures as paper surveys are burdensome and expensive; in both time and resources (Massagli and Carline, 2007). Compounding the administrative workload is the time and effort needed to secure an adequate number of assessors from multiple professional groups and to, finally, collate data upon collection. As well, recall bias, or the ability to incorrectly recall past experiences often more frequent in those with less exposure, must be accounted for (Parfrey and Barrett, 2009). To prevent recall bias from skewing results evaluators are often pressed with very short deadlines to complete the assessment.

Secondly, as mentioned, Sargeant *et al* (2005) suggest the potential benefits of MSF may be impaired by emotional reactions to negative evaluations, which may either cause a loss in participation or positively skewed results by raters. Without confidentiality, many raters may decline to participate. Joshi et al (2004) found that nurses were more enthusiastic about participation when informed of confidentiality in their assessments.

Nowack (2011) summarized studies in neurology suggesting that interpersonal stress – such as being judged by, or compared to, others – caused increased physiological stress response which lasted 50% longer than compared to individual judgement. One study from the business industry demonstrated that favorable comments were associated with improved performance (Smith and Walker, 2004). As well, individuals who received less unfavorable feedback showed greater improvement and, conversely, individuals who received more unfavorable feedback declined in performance (Smith and Walker, 2004).

There appears to be no absolute minimum number of evaluations necessary to adequately assess medical residents via MSF. Wood et al. (2004), in summarizing four previous studies, found the literature suggested a large range (from 20 to >50) of unique assessments were needed to produce reliable results. Thus, too few raters could limit reliability of rating where as requiring too many raters would prove difficult for recruitment (Lockyer et al, 2003). However, this is contradicted by other studies indicating as few as 4 raters could provide adequate inter-rater agreement (Overeem et al, 2012)

2.5.3 Nurses and Allied Health Professionals as Part of the Assessment Team

Implementing MSF for reliable evaluation purposes requires precisely what the term describes: multiple sources. This is especially important in evaluation of residents as they are frequently interacting with a wide array of healthcare professionals – as well as patients and families - ranging from, but not inclusive to, attending physicians, fellow residents, undergraduate medical students, register nurses (RNs), licensed practical nurses

(LPNs), physiotherapists, occupational therapists, social workers, home care workers, respiratory therapists pharmacists, speech language pathologists, dieticians, pastoral care, and facility staff.

As mentioned, this thesis distributes raters into 3 groups: physicians, nurses, and allied health professionals. Unfortunately, nursing staff are often infrequently involved in resident evaluation as commonly it is only the attending physicians participating in completing surveys or questionnaires for a specific rotation (Johnson & Cujec, 1989). One potential reason for the infrequent involvement of nurses in resident assessment is that residents may have difficulty accepting nurses' capability to accurately assess performance (Rezler, 1986). Regardless, nurses are vital members of the circle of care for every patient. They have the unique opportunity to interact with residents as well as witness their day-to-day behaviors and actions. Nursing staff may observe different aspects - such as team relationships, interactions with patients and family, and humanistic attitudes - of a resident's performance that may not be viewed by attending physicians and thus may offer a unique perspective during resident assessment (Johnson & Cujec, 1989; Risucci, 1989). The ability of the residents to create and maintain positive collaborative relationships with nursing staff is essential for patient safety and establishing a mutually supportive clinical environment (Ogunyemi et al, 2009).

Although the frequency and depth of interactions with each of nurse and allied health professionals will vary, due to a multitude of factors, they should *all* be considered potential members of an evaluation team. However, the literature appears to be divided as

to whether or not raters from varying health professions can show agreement when assessing residents on their humanistic qualities.

Two studies in particular demonstrate support for involving multiple perspectives in the MSF process for resident evaluation. First, Massagli and Carline (2007) found that physicians, faculty, nurses, and patients can reliably rate physicians' humanistic behavior. Their study consisted of three rater groups: nurses, allied health professionals (as defined above), and medical students. Over a three year period, after a physical medicine and rehabilitation rotation with one of the 28 participating residents, raters used a 5-point scale (1 = poor, 5 = outstanding) to rate residents' humanistic qualities over 12-statements. A total of 930 ratings were submitted consisting of 60% (n = 556) from allied health, 22% (n = 206) from nurses, and 18% (n = 168) from medical students. Analysis revealed inter-rater correlation of 0.77 to 0.90, where >0.7 is the standard for reliability. No mean ratings per group were provided by authors as they were focused on determining reliability of the tool and not investigating ratings in general.

A second supportive study was published by Joshi et al in 2004 involving the assessment of interpersonal and communication skills of eight obstetrics and gynecology residents. A 10-item, 5-point frequency scale (1 = "never", 5 = "always") questionnaire was distributed to nurses, faculty members, fellow residents, allied health professionals, medical students, patients, and a self-assessment by the resident. Intraclass correlations (agreement within a particular group) were strong ranging from 0.54 (patients) to 0.85(nurses). The authors rank ordered each resident by the mean of the rater group – i.e.

when nurses rated a resident low or high, so did the other groups – to demonstrate strong reliability between groups, such as $r = 0.81$ ($p = 0.016$) between medical students and fellow residents.

In contrast, two studies are presented to demonstrate the lack of agreement and reliability between rater groups using MSF. Weigelt et al published a 2004 paper entitled *The 360-Degree Evaluation: Increased Work with Little Return*. Ten trauma and surgical intensive care residents were assessed on how they performed regarding the six ACGME roles through the use of a 23-item, 9-point performance scale questionnaire (1 = "worst performance" and 9 = "best performance"). Rater groups consisted of chief resident, faculty, nurses (surgical intensive care unit, trauma, and nurse practitioner), administrative staff, and a resident self-assessment. Results found similarities within rater groups but no statistically significant correlations between rater groups. Interestingly, nurses rated residents lowest in areas of patient care and professionalism. A second contrasting study was conducted by Johnson & Cujec in 1989. Three attending physicians and six nurses independently assessed professional attributes, technical skills, knowledge, and overall competence of 60 residents from various specialities (surgical, medical, anesthesia, and obstetric residents) at the University of Saskatchewan who rotated through the ICU over a two month period. Residents also provided a self-assessment. Results illustrated agreement, via correlation, between physicians and nurses in all categories *except* on humanistic qualities in the resident. The authors defined humanistic qualities as including: integrity, respect, compassion, empathy, sensitivity, tolerance,

patient-centric, providing comfort and encouragement, reliability, trust, good rapport with staff and families.

2.6 Gender Bias

A secondary goal in our study included investigating whether residents were scored differently based on the gender of their raters. The idea of gender bias in the employment world is not a new concept. Much progress has been made over the last few decades to have males and females treated equally by superiors and co-workers. The medical profession has, and is, not immune to such bias but evidence suggests it is moving in the right direction. One example of this progress comes from an analysis of data from Canadian Residency Matching Service (CaRMS) for the 2012 residency selection period which demonstrated that slightly more women were selected for their desired residency position than males were (CaRMS, 2012). At the Canadian undergraduate level, a 2010 CaRMS study reported that the number female medical students heavily outweighed male medical students (58.2% vs. 41.8%) (CaRMS, 2010). In the United States, a database maintained by the Association of American Medical Colleges (AAMC) and American College and Surgeons (ACS) was analyzed for a similar study. The results demonstrated that the number of female applicants to various residency programs has been increasing in recent history (Davis *et al*, 2011). The analysis compared the gender of post-graduate applicants over six years (2000 – 2006) and across eight medical specialities. The results found that seven of the eight specialities had an increase in female applicants over the six year period, ranging from 3% (plastic surgery) to 12% (urology). At the undergraduate

medical level, the AAMC's public database indicates that female applicants and entering students, since 2000, have comprised, on average, 47% of the respective populations.

Regardless, biases still exist at all levels in the medical world. Trix and Psenka (2003) provide such an example in the analysis of 300 letters of recommendation for medical faculty positions. Letters of recommendation are a mandatory and crucial portion of any individual's application to attain the next professional level in their career. The study revealed biases in the quality of letters of recommendation between males and females. Points of disparity disadvantaging females include: shorter letters, more negative language, less demonstration of having a connection or relationship with applicant, and less promotion of potential skills and abilities (Trix & Psenka, 2003). A similar gender bias investigation was performed as a randomized double-blind study at Yale University in 2012 (Moss-Racusin, Dovidio, Brescoll, Graham, and Handelsman, 2012). Science faculty members individually rated an application of a single student – identical in content and randomly assigned as male, John (n = 63) or female, Jill (n = 64) - applying for a laboratory manager position. Results found that participants rated male applicants significantly more competent and hireable than the identical female applicant as well as offering a higher starting salary with more mentorship training (Moss-Racusin *et al*, 2012).

Investigating whether the gender of a medical resident and/or their raters would have an impact on the score received during a performance assessment became a secondary goal in this study considering the previous examples as well as local anecdotal evidence.

Informal conversations from residents suggested that female residents feel more negatively critiqued than their male counterparts; particularly by female nurses. A review of the literature on this topic found 13 articles with a wide array of seemingly contradictory results. Five studies compared scores received by male and female residents on a variety of educational dimensions or at licensing examinations without investigating interactions between rater and rate genders (i.e., female assessors providing lower scores to female students compared to male students). Three of those studies demonstrated that gender was a statistically significant factor in the score received – females scoring higher in males in two of the studies (Smith et al, 1991; Day et al, 1989, Ferguson et al, 2002) – while the two other studies demonstrated no significant gender difference (Massagli & Carline, 2007; Campbell et al, 2012).

The remaining eight studies investigated potential gender bias in assessment by analyzing the interaction between rater and ratee gender on overall scores. The existence of an interacting gender bias was split as four of the eight studies reported bias, four found no evidence of bias. Of those reporting a gender bias, the earliest study identified was by Kaplan and Centor in 1990. The authors found that female physicians received significantly more favorable evaluations than did male physicians when evaluated by female nurses – contradictory to our local anecdotal evidence. Rand *et al* (1998) found that male residents were graded significantly higher by male attending physicians relative to female attending physicians in six of nine assessment dimensions ($p < 0.01$) as well as on overall, or global, scoring, $p < 0.01$. An interesting, although statistically non-significant, trend indicated that female residents tended to receive higher evaluation

scores from female attending physicians than male attending physicians in eight of nine dimensions, no p-value provided (Rand *et al*, 1998). Wiskin *et al* (2004) found that female students performed better than males on 10 of 11 scenarios in objective standardized clinical examinations (OSCE), $p = 0.017$. As well, male examiners were found to provide lower scores than female examiners, $p = 0.043$. Another linguistic analysis conducted by Isaac *et al* (2011) provided further insight into gender bias during medical student assessment. The authors analyzed 297 letters from the Dean for undergraduate medicals students applying to a specific radiology residency program. Significant effects indicated that female authors of male students' recommendations used the fewest positive emotion words compared to all other gender pairing, $p = 0.006$. As well, female authors of female students used more positive verbs than for male students, $p = 0.027$.

In direct disagreement are four studies which determined gender bias did not exist in the assessment of medical students. Colliver *et al* (1993) analyzed non-clinical scores of fourth year medical students provided by standardized-patients during a mandatory OSCE over four years (1988 to 1991). The analysis found no interaction between rater and student gender over any of the years, p-value range = 0.165 – 0.735. Using a randomized, controlled design, Brienza *et al* (2004) found no statistically significant difference in the ratings of 160 residents (PGY 1- 3) by 88 faculty at Yale University, p-value range = 0.07 – 0.84. An intriguing, although statistically non-significant, trend indicated that male faculty ratings of female residents were lower than other pairings, $p = 0.07$. A third study by Thackeray *et al* (2012) specifically investigated gender bias in assessment in a

gastroenterology residency program. The results from 240 resident assessments by 44 faculty physicians found no statistical difference between resident gender, faculty gender, or an interaction between resident and faculty gender, p-value range = 0.24 – 0.72). Finally, strong evidence to support a lack of gender bias in medical education was provided by Dorsey and Colliver in 1995. Due to concerns of an existing gender bias, the Southern Illinois University School of Medicine changed their policy on examination assessment. The new policy ensured that student identifying information would be removed from exams before assessment. Eight years of data was collected for pre- and post-policy changes, four years of data for each group. The results found no difference between pre- and post-policy change groups regarding gender. A final, interesting, statistic from a Sax et al (2003) study investigating response rates found that females were twice as likely to respond to surveys than males (26.6% vs. 13.4%).

2.7 Inter-Rater Reliability

The concept of reliability has been previously introduced as presenting in two main varieties: test-retest reliability and internal consistency reliability. One of the research goals of this project was to determine a similar, yet, slightly unique type of reliability, of the ICAR: *inter-rater reliability*. Determining the inter-rater reliability of a tool seems like a relatively straight forward task upon initial investigation. However, to accurately determine the true level of agreement between – or even within – rater groups, when assessing a particular subject (e.g., medical resident), requires intensive statistical analysis.

Regardless of whether the trait or characteristic in a subject can be assessed with pure objectivity or subjectively, the instrument used for analysis must prove to be reliable, or provide consistent measurements over multiple assessments. Ideally, a perfectly reliable tool allows exact agreement between every assessment from any, and all, raters. Alas, the old adage 'to err is to be human' holds true and provides the starting point for this discussion.

The earliest model of reliability is termed 'classical test theory' which simply states that a score obtained on any given assessment consists of two parts: a true, or universal, score - the *actual* measure of the investigated trait - and some degree of error provided by sources of variance (McMillan & Schumacher, 2006).

$$\text{Observed Score} = \text{True / Universe Score} + \text{Error}$$

Sources of error can range from general factors, such as ambiguity in item wording, to specific to the measurement itself, such as a slightly ill-calibrated tool (ex. body mass scale). The more error, or sources of error, discovered the less reliable repeated measures will become, and vice versa. The following discussion outlines an approach to determining a tool's reliability - both internally and externally, such as with inter-rater reliability – and where its sources of variance and error reside and can be accounted for.

Once a desired instrument is developed with appropriate construct and content validity it is often utilized in a pilot study to determine its reliability before implementation into a large study design. Upon data collection and entry into a statistical software package, the

first step in a reliability analysis is to determine its *internal consistency*, that is, how well does the different items on the same tool appear to be measuring the same desired trait (Streiner and Norman, 1987). Cronbach's alpha, as mentioned, is one of the appropriate statistics for this test, as it measures the level of consistency, from 0 to 1, in rater's scores across multiple questions regarding a specific trait. For example, the ICAR has six domains, or constructs, measuring unique aspects of a general characteristic: collaboration. If rater scores remain consistent, say above 7, out of 9, across a domain then a high Cronbach's alpha will be calculated. Another practical view demonstrating a high level of consistency occurs when, across all items, 'rater A' gives a high score while 'rater B' scores low. Interpretively, a Cronbach's alpha value larger than 0.7 is generally accepted as having satisfactory internal consistency (McMillan & Schumacher, 2006).

The statistical approach in determining inter-rater reliability is accompanied by several layers of complexity. The literature diverges on what methodology is most appropriate when attempting to investigate the level of agreement between two or more raters observing a subject performing or displaying a specific behavior or skill. There are several methods to tackle inter-rater reliability analyses. Four common approaches, each with its own pitfalls, include: percent agreement, Cohen's Kappa and Fleiss' Kappa statistics, and Generalizability Theory.

The simplest, and least specific, method in calculating inter-rater reliability is by calculating the percent agreement between raters. Percent agreement is commonly used in inter-rater reliability studies over more complex analysis likely because it is easy to

compute and comprehend (Hayes & Hatch, 1999). Percent agreement is considered the most basic form of inter-rater reliability due to its lack of accounting for chance agreement, thus inflating the true agreement between raters. Hayes and Hatch (1999) caution using percent agreement in research as it is possible to agree, by chance, even if one rater does not observe the same subject interaction as another rater. The formula for percent agreement, as Figure 3 illustrates, calculates the number of exact agreements between raters divided by the total possible agreements.

$$\text{Percent agreement} = \frac{\text{Observed Agreement}}{\text{Possible Total Agreement}} \times 100\%$$

Figure 3 - Formula for Percent Agreement

To avoid the consequences of percent agreement, correlation measures such as a Kappa statistic are commonly reported in literature. Although there are many derivatives of the Kappa statistic, Cohen's Kappa (κ) is the most basic, statistically speaking, and is used to determine agreement between two raters when using nominal or ordinal data. Cohen's Kappa compares the exact agreement between raters while accounting for chance agreement, as seen in Figure 4, allowing for an increasingly true measure (Cohen, 1960).

$$\text{Cohen's Kappa } (\kappa) = \frac{\text{Observed agreement} - \text{Chance agreement}}{\text{Possible total agreement} - \text{Chance agreement}}$$

Figure 4 – General Formula for Calculating Cohen's Kappa (κ)

As it is a correlation coefficient, Cohen's Kappa is scored between 0 and 1, where 0 implies agreement is equal to chance and 1implies perfect agreement. Values less than 0

are possible if raters are agreeing less than chance would allow. Table 1 illustrates the qualitative level of agreement when interpreting a calculated Cohen's Kappa statistic.

Table 1 – Interpretation of Cohen's Kappa (κ) Statistic

Coefficient	**Interpretation**
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Landis & Koch (1977)

A caveat of Cohen's Kappa is that it only allows statistical analysis between two raters, which is not always the case in evaluation procedures. When comparing agreement between two or more raters, a similar analysis to Cohen's Kappa is used in *Fleiss' Kappa.* Fleiss' Kappa is an agreement coefficient that provides the level of agreement above chance – interpreted using the same 0 to 1 scale as Cohen's Kappa. Figure 7 illustrates the generic statistical formula for calculating Fleiss' Kappa statistic.

$$\text{Fleiss' Kappa} = \frac{\text{Observed agreement} - \text{Chance agreement}}{1 - \text{Chance agreement}}$$

Figure 7 - General Formula for Calculating Fleiss' Kappa

Finally, the most thorough method to investigate inter-rater reliability is through employing Generalizability Theory, or G-theory. In essence, G-theory is an extension of classical test theory where it not only provides an estimate of error but determines the sources of error through multiple analysis of variance (ANOVA) calculations. G-theory, developed by Cronbach et al (1972), attempts to account for all possible sources of

variance – termed *facets* – in a measurement and determine the extent to which these sources influence the result. For example, residents may consistently score higher on one section of an assessment. It is important to determine the level of variability this skewed result has on the residents' overall score. ANOVA calculations will provide evidence of the influence, or effect size, that each variable has on the true score. If the final residual, or error, term is large, then it can be assumed that there are more sources of variance than the facets in our model. A major advantage of G-theory is due to the fact that multiple tests for reliability (e.g., intra-observer and test-retest) can be incorporated into the same analysis instead of two separate analyses (Streiner and Norman, 1989). This allows for more precise results as well as variance estimates due to facet interaction. As with most reliability analyses, a generalizability coefficient is calculated using the mean sum of squares, from the ANOVA, of main facet and facet interaction.

Major challenges to using generalizability theory stem from the lack of available computer software and the rigorous effort needed to collect enough data for each facet to accurate calculate estimates of error variance (Alkharusi, 2012).

The literature review conducted provides evidence for the necessity of developing and implementing a new, appropriate, reliable, and valid assessment tool for the CanMEDS role of Collaborator.

Chapter 3 Methodology

3.1 Phase I – Pilot study of ICAR Reliability in Anesthesia

3.1.1 Goals

The primary goal of the first phase, pilot study was to determine the inter-rater reliability of the ICAR on a smaller-scale before introducing it to a larger multi-source feedback field test. Secondary objectives of this phase were two-fold: first, investigate the level of participation and general buy-in from attending physicians and residents regarding daily assessments using the ICAR; second, determine face validity of the ICAR and revise the tool if necessary.

3.1.2 ICAR Revision

When this project was initially proposed to Memorial University's Anesthesia residency program (Faculty of Medicine), four Anesthesia faculty physicians were asked to assess the 31-item ICAR for face validity with respect to assessment of their residents. Their feedback provided information regarding which items pertained to their program and would likely be interpreted correctly by the raters (e.g., was the educational language used on the ICAR understandable by physicians). Feedback from the four Anesthesia faculty physicians was compared by determining the level of agreement on each item. An item was removed or retained if three physicians agreed. If their vote was split, the item was retained. The analysis pared the initial 31-item ICAR down to 17 items. The document was also reduced from seven pages to four pages (Appendix C).

3.1.3 Description and Data Collection

The Anesthesia residency program was chosen for this study for two main reasons. First, in Anesthesia, attending physician – resident interactions were considered to be more 'controlled' than in many other disciplines. Anesthesia residents are rotated daily among different attending physicians. Much of the physician – resident interaction in a day is likely to be spent in the same environment – i.e. in the operating room. This quality of interaction should provide a valid assessment of a resident's collaboration, among other, skills. Secondly, with an accreditation process from the RCPSC in the near future there was an increased motivation for high participation from both attending physicians and residents.

Ethical approval was provided by Memorial University's Interdisciplinary Committee on Ethics in Human Research (ICEHR). Informed consent was necessary for all physicians and residents to participate in the study. Separate meetings were held with anesthesiology residents and attending physicians to discuss the study's purpose and design, their role in the study, and distribute informed consent forms. As well, these meetings provided feedback on potential issues that either group may see arising. Aside from specific meetings, study information was distributed to all potential participants via e-mail and on a monthly Anesthesia departmental newsletter. The intention was that roles, responsibilities and daily interactions of attending physicians and residents should not change or be impacted by this study.

Assessments were based on the attending physician – resident interactions that occurred during a single shift. Throughout the two week study period, each participating resident was assessed by as many attending physicians as possible. ICAR forms were distributed to participating attending physicians every morning before the start of their shift. An attending physician received an ICAR if they were assigned a participating resident or had not previously assessed that specific resident. Completed ICAR forms were collected and submitted to a blinded data analyst at the end of each day. To ensure blinding, attending physicians and residents were coded with a letter – 'A' or 'R', respectively – and a number (not in accordance with alphabetic order). For example, one coding combination for an assessment could be: 'A8 – R3'. The coding was assigned by the principal investigator and the master key was kept in both paper and digital format. The paper copy was stored in a secured filing cabinet in research supervisor's office while the digital format was stored on a password protected laptop.

3.1.4 Inclusion/Exclusion Criteria

All Anesthesia attending physicians and residents were invited to participate. There was no exclusion criteria enforced in the recruitment process as our sample size was limited due to the finite number of attending physicians and residents in the Anesthesia program.

3.1.5 Statistical Analysis

Data from all collected forms were inputted into SPSS/PASW version 19 as questionnaires were completed. The initial SPSS dataset included variables (columns) for rater code, resident code, the 17 ICAR items, and open text comments. Although no

residents were excluded from the recruitment process, they were excluded from data analysis if they did not have at least three assessments completed. Internal consistency was determined for the complete ICAR, as well as within each of the six domains (i.e. Communication, Collaboration, Roles and Responsibility, Collaborative Patient / Client – Family Centred, Team Functioning, Conflict Management / Resolution) using Cronbach's alpha; a value of > 0.70 indicates statistically significant internal consistency (Cronbach, 1951). An exploratory factor analysis using varimax rotation was conducted to determine whether or not the domains were independent subscales within the ICAR. Fleiss' Kappa statistic was used to assess inter-rater reliability between multiple raters with a value of > 0.70 indicating statistically significant agreement (Fleiss, 1971). AgreeStat Version 2011.3, a statistical software program, was purchased to calculate Fleiss' Kappa statistic.

3.1.6 Pilot Study Results

The pilot study resulted in 24 attending physicians completing at least one ICAR (60% of faculty) assessing a total of 11 participating residents (55% of residents). Of those 11 residents, only seven (64%) received at least 3 assessments (range, 3 – 7 raters per resident), and thus were included in the analysis.

Table 2 summarizes the internal consistency reliability analysis performed on the 4-point scale version of the ICAR as well as each of the six domains. The Cronbach's alpha value for the full ICAR instrument was 0.939. Cronbach's alpha values ranged from 0.667 (Roles and Responsibilities) to 0.876 (Collaboration).

Table 2: Pilot Study - Summary of ICAR and Domain Internal Consistency

Competency Domain	Cronbach's Alpha
Communication (4 items)	.768*
Collaboration (3 items)	.876*
Roles and Responsibility (3 items)	.667
Collaborative Patient/Client – Family Centred (2 items)	.800*
Team Functioning (2 items)	.708*
Conflict Management / Resolution (2 items)	.851*
ICAR (17 items)	**.939***

* > .70 indicates acceptable reliability

Fleiss' Kappa statistic calculated an inter-rater reliability coefficient of .003 (95% CI, .000 – .038) indicating no agreement between raters more than chance would allow (Landis & Koch, 1977). The percent agreement between raters was 66.8% (95% CI, 64.5% – 69.2%) across 31 raters over 17 items, adjusted for missing data.

Of the 527 total observations in the pilot study, 69 (13.1%) were deemed missing or non-observable, ranging from 0%, on the first item – measuring communication – to 54.8%, on the final item – measuring conflict management / resolution. Specifically, the conflict management / resolution domain, the final three items, averaged 33.3% missing or non-observable data.

Chapter 4 Phase 2 – Multi-Source Feedback Study

4.1 Methodology

4.1.1 Goals

The primary goal of the MSF field test was to assess the inter-rater reliability of the ICAR. Secondary outcomes investigated whether independent variables – demographic characteristics of raters – lead to a significant difference in overall ICAR score or global score or predict achievement of a specific score or higher.

4.1.2 ICAR Revision

As section 4.1indicates, the statistical analysis of inter-rater reliability from the pilot study revealed a poor level of agreement (Fleiss' Kappa = .003, 95% CI .000 – .038). A Kappa value of ~ .000 indicates that rater agreement is equal to agreement by chance alone. We hypothesized that the resulting positive skew was likely due to the natural disposition of medical residents functioning as high-achieving individuals. As such, they were unlikely to receive poorer scores such as a 1 (minimal) or 2 (developing); as was seen from pilot study results. In essence, the ICAR became a 2-point scale which allows higher agreement by chance, thus limiting the Kappa statistic. Streiner and Norman (1978) cite simulation studies that found a 5-point scale reduces reliability coefficients, such as Kappa, by up to 12% compared to using a 9-point scale. A 2009 study by Cook and Beckman investigated the difference in accuracy and inter-rater reliability between 5-point and 9-point scales in medical education assessments. The results found that 9-point

scales were more accurate than 5-point scales (54% v 44%, $p < .0001$) but no significant difference in inter-rater reliability was found between the two formats.

Due to our pilot study result as well as the cited literature, we expanded the ICAR scale from a 4-point scale to a 9-point scale where 1 = well below expectations, 5 = meets expectations, and 9 = well above expectations (Appendix D).

4.1.3 Description and Data Collection

Four residency programs (Internal Medicine, Neurology, Obstetrics / Gynecology, and Orthopedic Surgery) at Memorial University's Faculty of Medicine were recruited to participate in the field test study. Residents, from these four disciplines, completed 4-week rotations on one of five medical / surgical units (Internal Medicine, Neurology, Obstetrics / Gynecology, Orthopedic Surgery and Intensive Care). Contact with nurses and allied health professionals were initiated upon correspondence with division managers of the participating medical / surgical units. Meetings were held with residents, nurses, and allied health professionals to explain the goals of, and their role in the study, as well as to provide informed consent forms. Attending physicians were recruited individually after a resident's rotation due to uncertainty in confirming which physician specifically would be interacting with the resident. Upon obtaining consent, each resident's and rater's identity was converted to a unique letter and number code. For example, an intensive care nurse was coded as ICN1, an orthopedic attending physician was OP4, and an internal medicine resident became IMR2.

The 9-point scale ICAR was presented as a 3-page document in landscape orientation (Appendix D). The first page consisted of a brief message regarding the study as well as six questions investigating demographic characteristics of raters. The remaining two pages consisted of the 17 evaluative items, a global rating statement, and two open-response questions. The top right corner of each page was denoted with the rater-resident pair. For example, OP1 – OR3 for an orthopedic attending physician assessing an orthopedic resident. Descriptive characteristics of raters investigated included: profession, gender, number of years of experience in profession (greater than or less than ten years), number of years of experience in current medical / surgical unit (greater than or less than ten years), frequency of interaction (greater or less than once per shift), and type of interaction (direct, indirect, or both). A direct interaction was defined as a face-to-face or phone conversation. An indirect interaction was defined as contact through chart notes, orders, or requests; discharge planning; hearing from other colleagues; or hearing from patient or family.

Division managers provided names and shift schedules for participating nurses and allied health professionals, thus allowing the ICAR to be prospectively marked with the rater-resident pair. When the assessment period commenced each rater was met with individually by the lead researcher to explain the study, provide informed consent forms (if not already signed), and distribute the ICAR. To ensure the rater assessed the correct resident, a detachable Post-It© note listing the resident's name was attached to each ICAR. Each rater was asked to be complete assessment within 24 hours and place in a

large envelope located in their unit. The completed assessments were collected twice per day at the beginning of each nursing shift.

Sixteen – four per discipline – randomly-selected residents were chosen from all participating residents to be assessed by eligible attending physicians, nurses and allied professionals. In total, six residents were assessed by at least two unique individuals from each rater group over the allotted data collection period and were incorporated in the statistical analysis. The six residents were considered representative of the resident population as they comprised at least four different medical disciplines covering each of the post-graduate years (PGY1 – 5). Residents were blind to which rotation they would be evaluated on. As well, residents were not aware of which specific healthcare professionals were assessing them. They were under the assumption that all health professionals they interacted with were part of the evaluation team.

4.1.4 Inclusion/Exclusion Criteria

Residents were excluded from the study if, during the assessment period, they were on a rotation, or elective, outside of the Health Sciences Centre in St. John's, NL, Canada. Residents were also excluded from the data analysis if they did not have at least two completed ICAR assessments from each rater group. Residents and program directors confirmed which attending physicians had an acceptable level of interaction with resident to complete an assessment. Nurses and allied health professionals were excluded if they missed one, or more, of the four weeks of the resident's rotation or felt as if they did not interact with the specific resident.

4.1.5 Statistical Analysis

Data from all collected forms were inputted into SPSS/PASW version 19 as questionnaires were completed. The initial SPSS dataset included variables (columns) for rater code, resident code, each demographic characteristic, the 17 ICAR statements, global score, and open text comments. Subsequent variables were created within the same dataset as required for specific analyses, such as the creation of a binary variable for logistic regression.

Missing data for all quantitative variables was replaced using a single imputation stochastic regression method (Enders, 2010). The stochastic regression imputation replaces a single missing data point by accounting for the mean of the case (row), specific variable (column), and grand mean (entire data set). Single imputation methods are often viewed as weaker methods when compared to advanced techniques as they create biased parameter estimates (Gold and Bentler, 2000). Advanced techniques dealing with missing data incude: maximum likelihood imputation – the most sophisticated – and multiple imputation – the creation of multiple data points for a single missing value (Enders, 2010). However, stochastic regression imputation is the only single imputation method with merit due to the addition of a residual, or error term, which accounts for variability lost by other traditional single imputation methods (Enders, 2010). This residual term is created by multiplying the variable's standard deviation by a randomly generated value from a standard normal distribution created by SPSS as depicted in Figure 7.

New Value = Rater Mean + Variable Mean - Grand Mean + Residual*

*Residual = Variable standard deviation x Random normal distribution value

Figure 7 – Generic Formula for Stochastic Regression Imputation

The resulting filled-in variables were saved as new variables within the same dataset as the original data. The demographic characteristic variables were transformed into new binary variables to allow adequate sample sizes for statistical analysis. For example: Question # 3 which asks for the rater to indicate their total years of experience in their profession had six possible options ranging from 'less than one year' to 'greater than 20 years'. The original variable, *totalexp*, was transformed to a new binary variable, *tenyearexp*, indicating 'yes (1)' for 'greater than 10 years' and 'no (0)' for 'less than 10 years'. The same process was followed for questions #4 – 6.

Frequency of descriptive characteristics for each rater group and distribution of raters per group across residents were compared using Pearson's Chi-Square test. Comparison of overall ICAR score and mean global score were compared using one-way ANOVA. One-way ANOVA was also used to compare overall ICAR score and mean global score based on descriptive characteristics of raters and residents. Repeated-measures, two-way ANOVA were utilized to test for within-subject and between-subject main effects and interactions of independent variables combinations across the 17 items of the ICAR between residents.

4.2 MSF Study Results

Figure 7 illustrates the participation of residents in the MSF portion of the study which resulted in 27 residents initially completing an informed consent form. Of those, eight (29.7%) residents were deemed eligible to participate based on the rotation they were completing during the time of the study. The final statistical analysis included six (22.2%) residents assessed by at least two raters from each rater group.

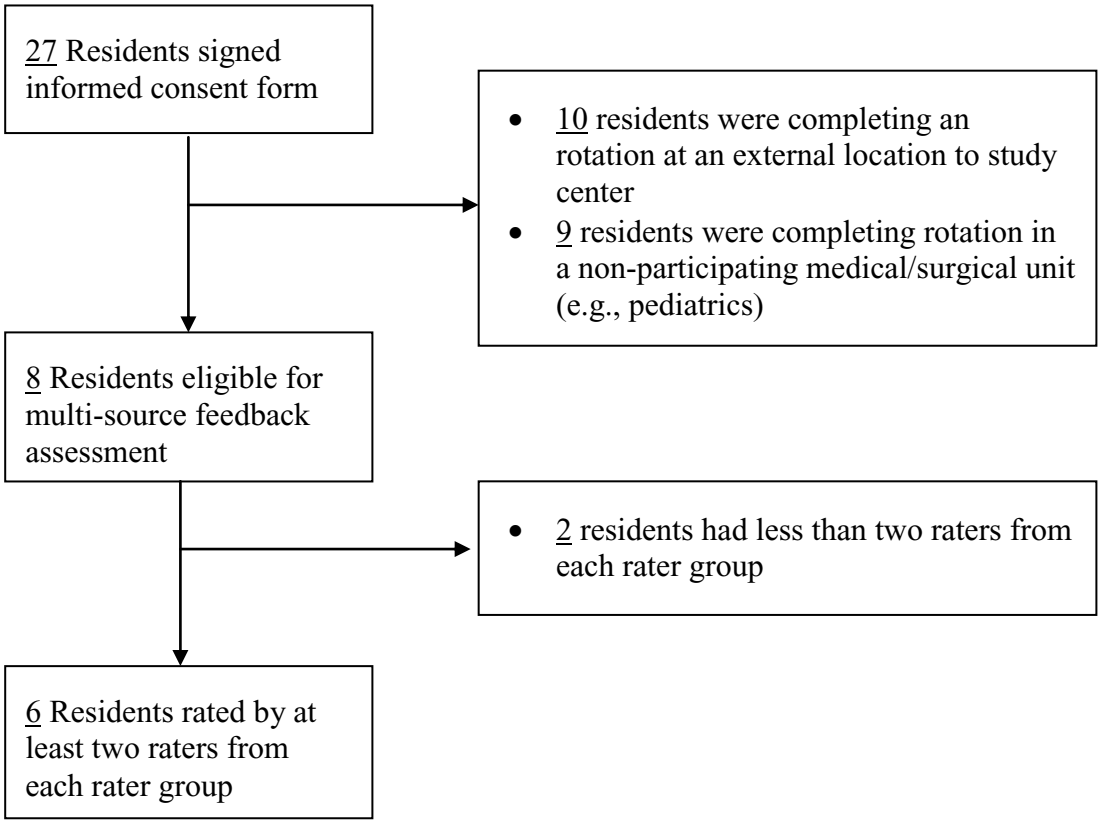


Figure 7: Flowchart of Resident Participation

4.2.1 Baseline Characteristics of Raters

Table 3 summarizes the frequency and proportion of demographic characteristics among rater groups. Nurses completed the majority of assessments (n = 107, 69.0%), followed by allied health professionals (n = 26, 16.8%), and physicians (n = 22, 14.2%). Females completed 81.3% (n = 126) of the total assessments. There were significant ($p < .001$) differences in the gender of participants from each rater group; male physicians (81.8%), female nurses (92.5%), and female allied health professionals (88.4%). There were more assessments completed by raters with at least 10 years of professional experience (60.0%) and in their current unit (55.5%). As well, the majority (65.8%) of assessments were completed by raters who reported at least one resident interaction per day.

Table 3 – Demographic Characteristics among Rater Groups

	Total	Physician	Nurse	Allied Health	χ^2	*p*
Ratings (n, %)	155	22 (14.2)	107 (69.0)	26 (16.8)		
Gender						
Female (n, %)	126 (81.3)	4 (18.2)	99 (92.5)	23 (88.5)	67.3	<.001*
Male (n, %)	29 (18.7)	18 (81.8)	8 (7.5)	3 (11.5)		
Years in Profession						
<10 (n, %)	62 (40.0)	7 (31.8)	45 (42.1)	10 (38.5)	0.83	.660
10+ (n, %)	93 (60.0)	15 (68.2)	62 (57.9)	16 (61.5)		
Years in Current						
<10 (n, %)	69 (44.5)	6 (27.3)	58 (54.2)	22 (84.6)	16.1	<.001*
10+ (n, %)	86 (55.5)	16 (72.7)	49 (45.8)	4 (15.4)		
Interaction						
≥ 1 per shift (n, %)	102 (65.8)	15 (68.2)	80 (75.5)	7 (26.9)	22.1	<.001*
< 1 per shift (n, %)	52 (33.5)	7 (31.8)	26 (24.5)	19 (73.1)		

* Significant at $\alpha = 0.05$

4.2.2 Rater Participation and Distribution

Table 3 summarizes the participation rates for each rater group. Of the 105 participating raters, 80 completed an assessment (76.2% response rate). Nurses and allied health professions had near equal response rates of 75.0% (n = 57) and 75.2% (n = 13), respectively. Physicians had the highest response rate of 90.9% (n = 10). Only one physician did not complete an assessment. Analysis found there was no significant difference in response rates between rater groups ($\chi^2 = 0.19$, $df = 2$, $p = .909$).

Also, 155 assessments were completed indicating that each rater completed, on average, 1.94 (or ~2) assessments. However, for the remaining analyses, each assessment was considered to be independent of the specific rater.

Table 4 – Summary of Participation among Rater Groups

	Consented	Completed	Response Rate
Physicians	11	10	90.9%
Nurses	76	57	75.0%
Allied Health Professionals	18	13	75.2%
Total	**105**	**80**	**76.2%**

Table 4 summarizes the distribution of rater groups across residents. Analysis found there was no significant difference in proportion of raters per resident (χ^2 = 13.412, *df* = 10, *p* = .202) with number of per resident raters ranging from 19 – 37. The ranges within rater groups across residents were: 2 –5 physicians; 10 – 30 nurses, and 4 – 5 allied health professionals.

Table 5 – Chi-Square Analysis of Rater Distribution across Residents

	Residents							
	A	**B**	**C**	**D**	**E**	**F**		
Rater Group							**Total**	*p*
Physicians	3	5	5	5	2	2	**22**	.202*
Nurses	16	10	11	11	30	29	**107**	
Allied Health	4	4	4	4	5	5	**26**	
Total	**23**	**19**	**20**	**20**	**37**	**36**	**155**	

*χ^2 = 13.412, *df* = 10

Table 5 summarizes the internal consistency reliability analysis performed on the 9-point scale version of the ICAR as well as each of the six domains. The analysis found that internal consistency reliability, through Cronbach's alpha coefficient, increased for overall 9-point scale version (α = .981) compared to the 4-point scale version (α = .939). Similarly, each of the six domains had an increased Cronbach's alpha coefficient. Differences in domains between ICAR versions ranged from α = .056 - .232. Due to the high internal consistency of the domains, the overall ICAR scores used in further analysis were the sum of all 17 items from the six domains.

Table 6 – Comparison of Internal Consistency Reliability between Pilot Study and MSF ICAR Formats

Competency Domain	**Cronbach's Alpha**	
	Pilot	MSF‡
Communication (4 items)	.768*	.963*
Collaboration (3 items)	.876*	.950*
Roles and Responsibility (3 items)	.667	.899*
Collaborative Patient/Client – Family Centred (2 items)	.800*	.881*
Team Functioning (2 items)	.708*	.932*
Conflict Management / Resolution (3 items)	.851*	.907*
		.981*

* > .70 indicates acceptable reliability

4.2.3 Missing Data Analysis

Table 7 compares the amount of missing data between pilot and MSF studies. The paired samples t-test indicates a significant reduction in missing data in the MSF study compared to the pilot study, 8.8% vs. 13.1% respectively, $p = .032$. The final two items on the ICAR, items #16 and #17 – both under the *Conflict Management / Resolution* domain – were reported as the highest percent missing in both studies, averaging 22.3% and 40.6%, respectively. A subsequent analysis compared the frequency of missing data averaged over each rater profession. Of the 234 missing data points, allied health professionals averaged 2.8 missing data values per rater, followed by nurses 1.3, and physicians 1.0.

Missing data was not replaced in the pilot study, but in the MSF study, missing data was replaced using a stochastic regression imputation method. A comparison of the mean and standard deviation between original and calculated datasets found differences of -0.05 (6.30 vs. 6.25) and -0.04 (1.49 vs. 1.45), respectively.

Table 7 – Comparison of Missing Data between Pilot Study and Multi-Source Feedback (MSF) Field Test Ordered by Highest Proportion Missing in Pilot Study

Item #	Item Category (# in Category)	Pilot (%)	MSF (%)	Difference (%)
17	Conflict Management / Resolution (3)	54.8	26.5	- 28.3
16	Conflict Management / Resolution (2)	25.8	18.7	- 7.1
8	Roles and Responsibility (1)	19.4	16.8	- 2.6
10	Roles and Responsibility (3)	19.4	15.5	- 3.9
15	Conflict Management / Resolution (1)	19.4	8.4	- 11.0
12	Patient/Client – Family Centred (2)	16.1	18.7	**+ 2.6**
14	Team Functioning (2)	16.1	3.9	- 12.2
11	Patient/Client – Family Centred (1)	12.9	17.4	**+ 4.5**
9	Roles and Responsibility (2)	9.7	7.1	- 2.6
13	Team Functioning (1)	9.7	5.8	- 3.9
6	Collaboration (2)	6.5	3.2	- 3.3
2	Communication (2)	3.2	1.3	- 1.9
3	Communication (3)	3.2	2.3	- 0.9
5	Collaboration (1)	3.2	3.2	0
7	Collaboration (3)	3.2	1.3	- 1.9
1	Communication (1)	0	0.6	**+ 0.6**
4	Communication (4)	0	0	0
	Total Missing	**13.1**	**8.8**	**- 4.3***

* Significant at $\alpha = 0.05$ (Paired samples t-test)

4.2.4 Comparison of Overall ICAR Scores

Table 8 summarizes the ANOVA analysis of overall ICAR scores across the 17 items on the ICAR between various descriptive characteristics of the raters. The only significant comparison indicated that female raters (n = 126, $\bar{x}$ = 6.12, SD = 1.03) scored residents lower than male raters (n = 29, $\bar{x}$ = 6.92, SD = 1.33), p = .008 yielding a small effect size of η^2= .045. Specific to MSF, there were no significant differences between rater groups, p = .297.

Table 8 – Comparison[‡] of Mean ICAR Scores for Independent Variables

	ICAR Scores					
	N	**Overall[α,β]**	**SD**	**F**	*p*	η^2
Profession				1.225	.297	.016
Physician	22	6.64	1.13			
Nurse	107	6.21	1.34			
Allied Health	26	6.09	1.30			
Gender of Rater				7.184	.008*	.045
Female	126	6.12	1.03			
Male	29	6.82	1.33			
Years in Profession				0.949	.331	.006
<10	62	6.12	1.27			
10+	93	6.33	1.32			
Years in Current Unit				0.011	.917	.000
<10	86	6.24	1.29			
10+	69	6.26	1.33			
Interaction Frequency				0.310	.579	.002
≥ 1 per shift	102	6.30	1.35			
< 1 per shift	52	6.18	1.22			
Gender of Resident				0.013	.908	.000
Female	2	6.23	1.34			
Male	4	6.26	1.29			

[‡] One-Way ANOVA
[*] Significant at α = 0.05
[α] Overall ICAR score determined by summing total score divided by total number of raters
[β] ICAR scored on a 9-point scale

4.2.4 Comparison of Mean Global Score

Table 9 summarizes the comparisons of mean scores produced from the global scale between various descriptive characteristics of the raters. Three significant differences resulted yielding small effect sizes. Raters with greater than 10 years of total experience (n = 93, $\bar{x}$ = 5.85, SD = 1.29) scored residents lower than raters with less than 10 years of experience (n = 62, $\bar{x}$ = 6.52, SD = 1.32), p = .002, yield a small effect size of η^2 = 0.062. Second, raters with greater than 10 years of experience in the current unit (n = 69, $\bar{x}$ = 5.87, SD = 1.30) scored residents lower than raters with less than 10 years of experience (n = 86, $\bar{x}$ = 6.30, SD = 1.35), p = .048. Finally, male residents (n = 4, $\bar{x}$ = 5.76, SD = 1.25) received lower scores than female residents (n = 2, $\bar{x}$ = 6.19, SD = 1.35), p = .004. As with overall ICAR score, there were no significant differences between rater groups on the global scale, p= .364.

Table 9 –Comparison of Mean Global Scores for Independent Variables

	Global Scores					
	N	**Mean[α,β]**	**s**	**F**	*p*	**η^2**
Profession						
Physician	22	5.80	1.27	1.018	.364	.013
Nurse	107	6.11	1.34			
Allied Health	26	6.35	1.38			
Gender of Rater				2.366	.126	.015
Female	126	6.19	1.35			
Male	29	5.76	1.25			
Years in Profession				10.168	.002*	.062
<10	62	6.52	1.32			
10+	93	5.85	1.29			
Years in Current Unit				3.989	.048*	.025
<10	86	6.30	1.35			
10+	69	5.87	1.30			
Interaction Frequency				0.951	.331	.006
≥ 1 per shift	102	6.25	1.34			
< 1 per shift	52	6.03	1.34			
Gender of Resident				8.574	.004*	.053
Female	2	6.49	1.32			
Male	4	5.86	1.29			

‡ One-Way ANOVA
* Significant at α = 0.05
α Overall ICAR score determined by summing total score divided by total number of raters
β ICAR scored on a 9-point scale

4.2.5 Summary of Repeated Measures ANOVA Analysis

Two-way repeated measures ANOVA were used to investigate significant differences and effect size (η^2) across the mean score for each of the 17 items and two specific factors , or independent variable (e.g., rater gender and rater profession). Table 10 summarizes the two-way repeated measures ANOVA analysis *within* each rater profession across the six residents. The analysis revealed a significant, but small, interaction effect between the items and residents ($F = 1.378$, $p = .040$, $\eta^2 = .048$) indicating a difference in overall ICAR score across the six residents, with a small effect size constituting 4.8% of the variance. Secondly, there was a significant main effect regarding means of the individual 17 items ($F = 2.79$, $p = .002$) with a low η^2 value of 0.02, indicating a small effect size accounting for 2% of the total variance. Post-hoc comparisons found that means of item #1 and #16 were significantly different, $p = .048$. Specific to MSF, the analysis also found that rater groups did not differ in their scores across items as indicated by a non-significant interaction effect ($F = 0.807$, $p = .713$, $\eta^2 = .012$).

Table 10 - Summary of Two-way Repeated Measures ANOVA Analysis for Within-Subject Effect of Rater Profession and Resident across ICAR Items

	SS	df	MS	F**	*p*	η^2
Items	24.043	10.431	2.305	2.79	.002*	.020
Items x Resident	59.391	52.154	1.139	1.378	.040*	.048
Items x Profession	13.907	20.862	0.667	0.807	.713	.012
Items x Resident x Profession	100.377	104.309	0.962	1.165	.130	.078
Error	1180.542	1429.03	0.826			

* Significant at $\alpha = 0.05$

**Geisser-Greenhouse utilized as sphericity assumption was rejected.

Table 11 summarizes the two-way repeated measures ANOVA analysis *between* rater professions across the six residents irrespective of the items. The analysis revealed a significant interaction between rater group and individual residents with a large effect size contributing 19.4% of the variance ($F_{10,137} = 3.298$, $p = .001$, $\eta^2 = .194$). Interaction effects occur for several residents as shown in Figure 8. The analysis revealed no significant main effects between residents ($F_{5,10} = 1.587$, $p = .168$, $\eta^2 = .055$) or between rater groups ($F_{2,10} = 1.005$, $p = .369$, $\eta^2 = .014$).

Table 11 – Summary of Two-way Repeated Measures ANOVA Analysis for Between-Subject Effect of Rater Profession and Resident

	SS	df	MS	F	*p*	η^2
Resident	183.441	5	36.688	1.587	.168	.055
Profession	46.471	2	23.236	1.005	.369	.014
Resident x Profession	762.418	10	76.242	3.298	.001*	.194
Error	3167.375	137	23.12			

* Significant at α = 0.05

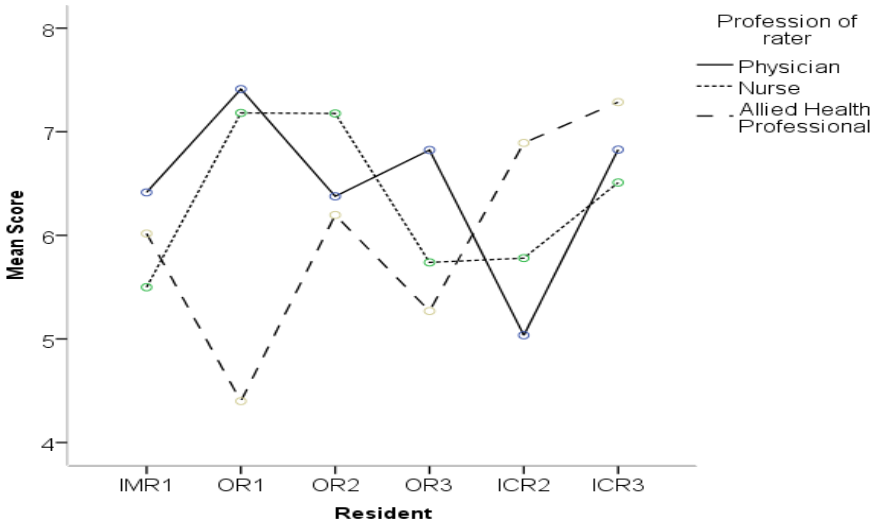


Figure 8: Interaction between Rater Profession and Residents

Table 12 and Figure 9 summarize the two-way repeated measures ANOVA analysis investigating whether the 17-item scores provided by the gender of the rater differ significantly across the six residents. The analysis revealed a significant interaction effect within rater gender across the means of the individual 17 items ($F = 1.911$, $p = .021$, $\eta^2 = .013$) accounting for only 1.3% of the total variance.

Table 12 – Summary of Two-way Repeated Measures ANOVA Analysis for Within-Subject Effect of Rater Genders and Resident across ICAR Items

	SS	df	MS	F	*p*	η^2
Items	29.043	10.623	2.734	3.368	.000*	.023
Items x Resident	62.247	53.113	1.172	1.444	.021*	.048
Items x Rater Gender	16.484	10.623	1.552	1.911	.036*	.013
Items x Resident x Rater Gender	47.282	53.113	0.89	1.097	.297	.037
Error	1233.236	1519.04	0.812			

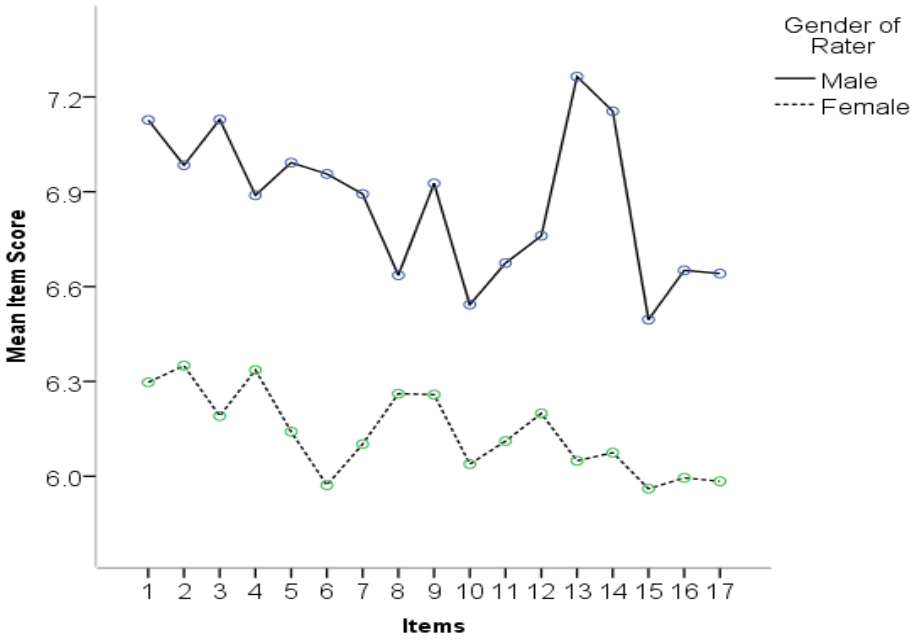


Figure 9: Interaction between Rater Genders and Items

Table 13 summarizes the two-way repeated measures ANOVA analysis *between* rater genders across the six residents irrespective of the items. The analysis revealed a significant main effect between rater gender ($F_{1,5} = 7.058$, $p = .009$, $\eta^2 = .047$) accounting for 4.7% of the total variance. The box plot in Figure 10 provides a visual comparison of the mean scores provided by each gender of rater.

Table 13 – Summary of Two-way Repeated Measures ANOVA Analysis for Between-Subject Effect of Rater Genders and Resident

Between Subjects	SS	df	MS	F	*p*	η^2
Intercept	57707.8	1	57707.8	2243.362	.000	.940
Resident	146.566	5	29.313	1.14	.342	.038
Rater Gender	181.546	1	181.546	7.058	.009*	.047
Resident x Rater Gender	142.495	5	28.499	1.108	.359	.037
Error	3678.504	143	25.724			

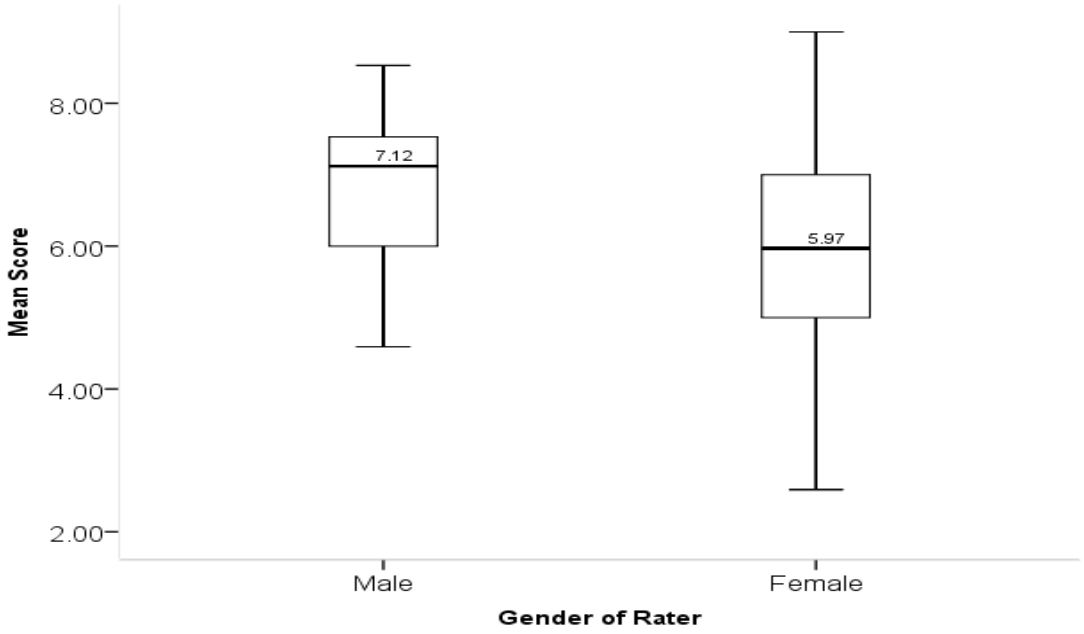


Figure 10: Box Plot of Mean Score Difference in Rater Genders

Table 14 summarizes the two-way repeated measures ANOVA analysis investigating whether the 17-item scores provided by rater interaction frequency differ significantly across six residents. The analysis revealed a significant interaction effect within interaction frequency groups across the means of the individual 17 items ($F = 2.103$, $p = .025$, $\eta^2 = .014$) accounting for only 1.4% of the total variance. Figure 11 clearly depicts items #5, #6, and #7 (all comprise the 'Collaborator' domain) being scored lower by raters who interact with residents less than once per shift.

Table 14 – Summary of Two-way Repeated Measures ANOVA Analysis for Within-Subject Effect of Interaction Frequency and Resident across ICAR Items

Within-Subjects	SS	df	MS	F	*p*	η^2
Items	32.476	10.837	2.997	3.752	.000*	.026
Items x Resident	60.205	54.185	1.111	1.391	.033*	.047
Items x Interaction Freq	17.426	10.837	1.608	2.013	.025*	.014
Items x Resident x Interaction Freq	48.079	54.185	0.887	1.111	.272	.038
Error	1229.241	1538.86	0.799			

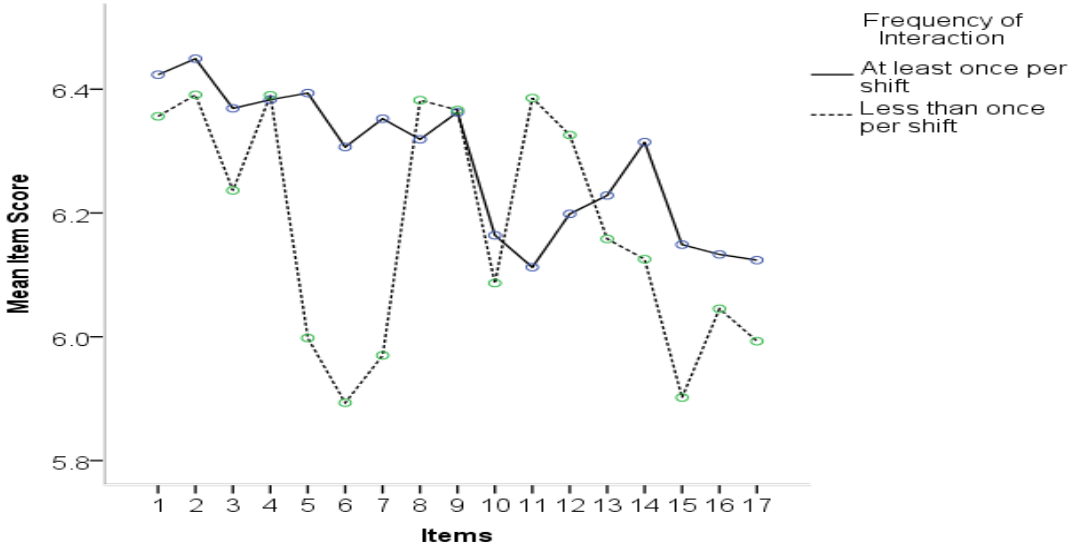


Figure 11: Interaction between Rater Interaction Frequency Groups and Items

Table 15 summarizes the two-way repeated measures ANOVA analysis *between* rater interaction frequency groups across the six residents irrespective of the items. The analysis revealed no main effect or significant difference between interaction frequency groups ($F_{1,5} = 0.224$, $p = .636$, $\eta^2 = .002$).

Table 15 – Summary of Two-way Repeated Measures ANOVA Analysis for Between-Subject Effect of Interaction Frequency and Resident

Between Subjects	**SS**	**df**	**MS**	**F**	***P***	η^2
Intercept	85034.298	1	85034.3	3186.126	.000	.957
Resident	296.068	5	59.214	2.219	.056	.072
Interaction Freq	5.987	1	5.987	0.224	.636	.002
Resident x Interaction Freq	180.894	5	36.179	1.356	.245	.046
Error	3789.828	142	26.689			

4.2.6 Logistic Regression

Logistic Regression was used to calculate the odds ratio for individual rater characteristics on a resident's likelihood to achieve an overall ICAR score and global score of above 6.0, 7.0, or 8.0.

Table 16 summarizes the logistic regression analysis of rater characteristics in predicting above a specific overall ICAR score. Rater gender was found to be the only significant predictor on overall ICAR score. Male raters were 3.08 times more likely than female raters to provide an overall ICAR score of above 6.0 (p = .013) and, more significantly, 3.28 times more likely to score above 7.0 (p = *.005*). The significance of rater gender did not exist for scoring above 8.0 (p = *.269*). Multi-variate logistic regression yielded no significant predictors with exception to rater gender.

Table 17 (page 77) summarizes the logistic regression analysis of rater characteristics in predicting above a specific global score. The most significant results found that male residents were 69.3% less likely to receive a global score above 7.0 (p = *.007*). The analysis also revealed significant odds ratios for year of experience. Raters with greater than ten years of both total years of experience and years current medical unit were less likely (56.4%, p = .015 and 52.0%, p = .034, respectively) to score above 6.0 than those raters with less than ten years of experience. Notable, slightly non-significant, results found that male raters were 61.6% less likely to score above 6.0. Multi-variate logistic regression yielded no significant predictors with exception to rater gender.

Table 16 – Summary of Univariate Logistic Regression Predicting Odds of Scoring Above a Specific Mean Score

	Mean Score > 6.0			Mean Score > 7.0			Mean Score > 8.0		
	(exp) β	95% CI	sig	(exp) β	95% CI	sig	(exp) β	95% CI	sig
Profession (vs. Physician)									
Nurse	.501	.194 – 1.29	.153	.468	.183 – 1.20	.113	.413	.294 – 19.7	.413
Allied Health	.779	.243 – 2.50	.675	.360	.104 – 1.24	.106	.657	.148 – 20.7	.657
Rater Gender (Male)	3.08	1.27 – 7.47	.013*	3.28	1.43 – 7.56	.005*	.310	.039 – 2.47	.269
Resident Gender (Male)	1.25	.658 – 2.39	.492	1.16	.567 – 2.35	.691	.451	.149 – 1.37	.161
Years in Profession (10+)	1.19	.624 – 2.26	.600	1.20	.590 – 2.44	.616	1.75	.522 – 5.84	.365
Years in Current Unit (10+)	1.09	.579 – 2.06	.788	1.07	.534 – 2.13	.853	.929	.306 – 2.82	.896
Interaction Frequency (> 1/shift)	.963	.494 – 1.88	.912	1.24	.591 – 2.61	.569	1.97	.526 – 7.41	.314

***significant at α=0.05**

Table 17 – Summary of Univariate Logistic Regression Predicting Odds of Scoring Above a Specific Global Score

	Global Score > 6.0			Global Score > 7.0			Global Score > 8.0		
	(exp) β	95% CI	sig	(exp) β	95% CI	sig	(exp) β	95% CI	sig
Profession (vs. Physician)									
Nurse	2.20	.754 – 6.41	.149	1.46	.392 – 5.40	.574	1.25	.143 – 10.9	.842
Allied Health	2.13	.595 – 7.58	.246	1.90	.415 – 8.70	.408	2.74	.264 – 28.4	.399
Rater Gender (Male)	.384	.146 – 1.01	.052	.444	.125 – 1.58	.210	1.09	.220 – 5.44	.914
Resident Gender (Male)	.526	.270 – 1.03	.059	.317	.137 - .730	.007*	.629	.174 – 2.27	.479
Years in Profession (10+)	.436	.223 – .853	.015*	.467	.206 – 1.06	.068	.419	.113 – 1.55	.193
Years in Current Unit (10+)	.480	.244 – .946	.034*	.568	.258 – 1.39	.231	.821	.222 – 3.03	.767
Interaction Frequency (> 1/shift)	.989	.494 – 1.98	.974	1.16	.489 – 2.78	.730	.313	.084 – 1.16	.083

***significant at α=0.05**

Chapter 5 Discussion

5.1 Pilot Study

The purpose of the initial pilot study was to determine the internal consistency reliability, inter-rater reliability, and the face validity of the Interprofessional Collaborator Assessment Rubric (ICAR). The ICAR demonstrated strong internal consistency reliability, through Cronbach's alpha, providing evidence of construct validity for the evaluative items. The internal consistency reliability for each of the six domains was also strong. Interestingly, a factor analysis found that all 17 items constituted a single construct: collaboration. Thus, the creation of six separate domains was unnecessary and should not be viewed as separate or distinct constructs within the ICAR. Analysis of missing and 'not observable' data revealed potential issues with some items on the ICAR. The final two evaluative items – #16 & #17, both in the *Conflict Management and Resolution* domain – were either 'non-observable' or missing in over one quarter (item #16) and one half (item #17) of completed assessments.

Reasons for such a high frequency of 'non-observable' or missing data is likely attributed to the nature of daily assessments. For instance, it's unlikely that there will be a conflict for a resident to deal with every day which would allow a rater to adequately assess a resident's skills in that area. Addressing this issue, one participating attending physician commented:

> "Many of these points are impossible to assess on a daily interaction. If you only have two patients and the resident performs a certain behavior is this sometimes? Frequently? Consistently? If we only have one patient, then does the behavior become always?"

Supporting the previous comment is another frequently raised issue by physicians. Depending on the resident's seniority, the attending physician's interaction with the resident may be limited. Senior residents are more likely to work independently, not under the guidance of their attending physician. This adds another factor to decreasing the likelihood of a rater appropriately assessing specific competencies. A comment from another participating attending physician echoes this observation:

> "Often we [attending physicians] do not observe the senior residents interactions with the patients. Also, the resident's interaction with nurses… are also not often observed by attending staff."

In addition to determining the internal consistency of the ICAR, we also investigated inter-rater reliability. Analysis found inter-rater agreement was equal to what chance alone would predict. This poor agreement value may be attributed to several factors. Residents were only assessed over a single shift, and not over an extended time interval, which may limit overall agreement as a resident's perceived collaboration skills may differ daily. As well, external factors such as stress, including individual coping skills and social supports may affect performance and inter-personal interactions on a day-to-day basis (LeBlanc, 2009).

The pilot study offered insightful qualitative – i.e., comments regarding the length of assessment time – and quantitative – i.e., skewed grading due to the 4-point scale – information that was incorporated in the design and implantation of the multi-source feedback field test which extended both assessment time and scale measurements.

5.2 Multi-Source Feedback (MSF)

5.2.1 Participation / response rates

One of our secondary outcomes investigated the general attitudes toward the acceptance of MSF as an evaluation tool in our medical faculty. The overall response rate (76.2%) in the field test of the modified ICAR in a MSF assessment process was generally high for all rater groups; ranging from 75.0% to 90.9%. This result reflects the upper end of response rates reported in the literature regarding MSF feedback which ranges from 36% (Hill et al., 2012) to 95% (Violato et al., 2008). This response rate suggests that the use of the ICAR in a MSF process with post-graduate residents may be a viable option to assess collaboration.

Despite the high participation rates for raters, residents had more negative feelings toward assessments provided from non-attending physicians. During recruitment meetings, we found that residents were hesitant to participate for a variety of reasons. First, as echoed by Rezler et al (1986), some residents questioned whether nurses or allied health professionals had the ability to evaluate resident performance. Residents noted, they, themselves, would find it difficult to evaluate nurses and allied health professionals on *their* abilities. Secondly, residents anticipated the quality of feedback received would

likely not be useful to their improvement. Such sentiments have been documented by Canavan et al (2010) during focus group discussions regarding MSF with residents. The authors found that residents received primarily positive feedback and a lack of suggestions for areas of improvement. Amin et al (2006) further support this finding, noting that a disadvantage of MSF involves raters hesitating to provide feedback to poorly performing learners.

5.2.2 Less missing data and distribution of missing data

The analysis of missing data (Table 7) for comparing the pilot study and MSF field test revealed interesting results. The statistically significant reduction of missing data between the pilot and the MSF study provides evidence that longer observation periods are more suitable for adequate assessment and evaluation of non-medical expert CanMEDS roles – those more subjective in nature – than daily assessments.

Both studies indicated that items in the *Conflict Resolution and Management* domain had the highest proportion of missing data. Respectfully and appropriately dealing with conflict is a critically important component of effective teamwork (Vivar, 2006). However, it appears to be difficult to assess – non-observable in approximately one-quarter of cases – due to a relatively low occurrence of observed conflict in an efficient medical unit. Also, as raters were asked to assess residents only over a 4-week time interval instead of a possible 4-month rotation, there would be less conflicts to be observed.

5.2.3 Agreement, mean score, and global score differences between rater groups

The primary outcome of this research project was to determine the level of inter-rater agreement between the three derived rater groups: attending physicians, nurses, and allied health professionals. Analysis of overall ICAR scores across all 17 items found no significant differences between rater groups. In juxtaposition, we discovered a low inter-rater reliability value between raters. But, as noted by Viera and Garrett (2005), a paradox may occur where a low Kappa statistic and high rater agreement co-exist. Initial hypotheses indicated that agreement of raters would add reliability and validity to the instrument and that the inclusion of other rater sources (nurses and allied health professionals), through a MSF approach, would provide a different viewpoint of the learner. However, our non-significant result – regarding overall resident score – quantitatively indicates that the additional rater sources may not observe anything different than the primary rater source: attending physicians. Interestingly, medical education literature overwhelmingly suggests MSF approaches to be used as formative assessments. The qualitative feedback (comments) from other rater sources is likely to be the most informative and useful to the learner.

The same non-significant finding was true regarding the analysis of the global scale, which asked raters to compare the resident being assessed to all previous residents they had previously interacted with. Interestingly, despite the non-significance, physicians scored residents highest with regards to overall ICAR score across the several measured

items but lowest regarding global score comparing the residents to all other residents that have interacted with the rater.

5.2.4 Gender Bias

A secondary outcome investigated the legitimacy of local resident's anecdotal perceptions of gender bias during their clinical rotations. Residents, particularly females, felt they had 'a harder time' than their male colleagues; particularly during interactions with female nurses. Our analysis revealed several significant results suggesting rater gender as main effect but not when the resident gender is accounted for; thus, providing evidence against the perceived gender myth.

As the ICAR utilized an expectation scale (below, meets, or above expectations), combined with the difficulty identifying the resident's true score given the subjective nature of Collaboration, it is impossible to determine if the statistically significant difference between rater gender indicates that female raters have higher expectations (i.e., score lower) or males have lower expectations (i.e., score higher) of resident collaborative ability. To this point, Ostroff et al (2004) investigated predictive ability of demographic, or descriptive, variables of raters on the score an individual would receive. Their analysis found that male raters tended to be over-estimators of an individual's performance.

Regardless, the significant difference between genders indicates that female raters will, on average, provide a learner with an overall lower score by 0.7 out of 9. Furthermore, our logistic regression analysis found that a learner was more than three times as likely to receive a score above the median if their rater was a male.

The initial idea for investigating possible gender bias in resident assessment was due to anecdotal evidence that female residents would receive poorer grades from female nurses than male residents. However, analysis of resident gender, regardless of rater profession, did not yield a significant difference in overall ICAR score. If anything, the majority of literature cites that female medical learners score higher than their fellow male students (Smith et al, 1991; Day et al, 1989; Kaplan and Centor, 1990; Rand *et al*, 1998; Wiskin *et al,* 2004). The final piece of evidence against potential gender bias was derived from the two-way repeated measures ANOVA investigating interactions between the gender of rater and resident. The resulting finding of no difference contributes to the wealth of pre-existing, and contradictory, gender bias literature.

Different results were found regarding gender bias using the global score measurement. Though, non-significant, male raters trended toward providing lower global scores than female raters. The logistic regression analysis found that resident gender *did* have a significant main effect. Male residents were more than twice as likely to receive a score below 7.0 (out of 9) than female residents.

The diverging results between the two measurement scales suggest that male raters have lower expectations over a short observation period but higher standards of residents overall than female raters. Such results are unfortunate as we assume equality exists in professional environments, especially with the assessments of future physicians. Such findings might imply that residents may not be receiving valid assessments throughout their residency training depending on the assessment procedures (i.e., assessed only by

attending physician) in place in their institution or program. Furthermore, some medical specialities, such as in the surgery fields, have a large male-to-female faculty ratio which could lead to residents in those specialities receiving inflated assessments.

5.2.5 Effect of other rater characteristics

No significant differences in overall ICAR score were found between raters when dichotomized into groups based on their total professional experience, experience in their current practicing unit, or their frequency of interaction with the resident. This is an important result as it addresses the resident's concerns on the ability of other rater groups to adequately assess them. For example, during initial meetings, one resident noted they were worried about assessment from nurses with a low level of experience. Another resident noted that they may not interact daily with allied health professionals.

Once again, the global score measurement provided contrasting results. Total and current unit experience of the rater provided a significant main effect in the global score received. Raters with more than ten years of experience provided lower global scores and were twice as likely to provide a score below the median than less experience raters. Although, no research could be found to support why senior staff may provide lower assessment scores, this result may be contributed to senior staff having higher expectations and standards for residents due to their experience interacting in both high-functioning and low-functioning team environments over their careers. Thus, they would likely be able to identify when a resident is providing a positive or negative addition during inter-professional interactions.

5.3 Strengths and Limitations

Major strengths of this study include:

The ICAR utilized in this study is previously validated assessment tool (Curran et al, 2011). Furthermore, the completion of a pilot study aided the iterative process in creating a more-refined assessment tool for the purpose of our main study.

Similar MSF studies have yielded response rates ranging between 36% and 95%. Our response rate of 76.2% was toward the higher end of studies. Furthermore, we were able to achieve a surprisingly high response rate, or buy-in, from physicians (90.9%). Great support and interest was also demonstrated by the Dean and Vice-Dean of Medicine, assistant dean of post-graduate medicine, as well as program directors for each of the participating residency programs.

Personal communication with the Newfoundland and Labrador (NL) Nurses Union indicated that the proportion of male nurses actively working in NL is 3.9% (238 males out of 6082 active nurses) whereas our study had 7.5% (8 males out of 107 participating nurses). This prevents any perception that there were too few male nurses in our study which would lead to skewed results. Unfortunately, in some research areas there is a natural bias which exists regarding the study population.

During the MSF phase of the study, each rater was met with individually to describe the study and the ICAR. As such, each rater was able to ask questions or voice concerns

regarding the study and their participation. Also, qualitative – written or verbal – feedback from the raters was positive and reaffirmed the need for such a study.

Although 8.8% missing data was missing upon collection from the MSF phase of the study, utilizing a stochastic regression imputation method allowed a full dataset to be analyzed without losing any significant data quality.

Three major limitations emerged from this study:

First, the study was conducted in a single institution on four medical units which may affect generalizability of results to other units, hospitals, regions, and provinces.

Second, regarding raters, there were low sample sizes of physicians and allied health (true ratios of the number of physicians / allied health professionals to nurses were not obtained). Furthermore, residents indicated which physicians were appropriate to assess them, thus it may have been possible that residents suggested raters who were likely to give more favorable results. Finally, no training was offered to raters regarding use of the ICAR as we wished to determine the feasibility and ease of use of the tool.

Third, regarding residents, there were a low number of residents eligible for inclusion into statistical analysis due difficulties obtaining adequate number of assessments from each rater group during the study period. Likewise, due to low numbers of residents overall, there was an uneven distribution of the gender of resident included in the analysis. Both limitations reduce the chance of uncovering a true effect.

Conclusion

Our research suggests that the ICAR, through a multi-source feedback process, would act as a strong formative assessment tool in graduate medical education.

Our conclusions regarding study outcomes:

1. **Inter-rater reliability of ICAR through multi-source feedback.**

   The results of the pilot study demonstrated a low inter-rater reliability coefficient of .003 (95% CI, .000 – .038) indicating no agreement between raters more than chance would allow (Landis & Koch, 1977). The percent agreement between raters was 66.8% (95% CI, 64.5% – 69.2%) across 31 raters over 17 items, adjusted for missing data. Several potential factors led to the low inter-rater reliability including the single-observation assessment and a 4-point rating scale. The low inter-rater reliability along with feedback from participants in the pilot project led to the changes in the ICAR format for the MSF study. The subsequent results from the MSF study demonstrated no significant differences in the ability of physician, nurse, and allied health professionals to assess medical residents in collaboration competencies.

2. **The feasibility of incorporating the ICAR? I.e., Will the ICAR be a tool that health professionals use to rate residents?**

   Based on the participation rates, where 75% - 90% of raters completed an assessment, it appears that the ICAR could be a viable tool used to assess post-graduate medical learner's collaboration competence. Furthermore, the ICAR was well-received by Dean of medicine and residency program directors.

3. **Resident's perceptions to the ICAR and MSF.**

   Although there was some initial hesitation from a minority of residents, there appeared to be an overall acceptance in using the ICAR to evaluate their collaboration ability. Residents need to be assured that nurses and allied health staff will be strictly assessing their collaboration skills, as determined by the CanMEDS Collaborator role, not other roles such as Medical Expert.

4. **Evaluation biases (gender, years of experience, etc.) exist when assessing collaboration in residents.**

   There were no significant differences in the overall ICAR score between three rater professions in the assessment of residents. Further analysis indicated that experience – overall or in their current area of work – of the rater and the frequency of interaction with the resident had no significant effect on the overall ICAR score.

   However, significant differences were discovered in overall ICAR score with regards to rater gender. Female raters provided a lower score (6.12 v. 6.82) than male raters regardless of the gender of the resident. Conversely, male residents scored significantly lower (5.76 v. 6.19) than female residents on the global rating scale.

   Also, regarding the global rating scale, raters with more than ten years of experience scored residents lower than raters with less than ten years of experience. Rater

profession, rater gender, and frequency of interaction with the resident had no significant effect on global rating score.

Future Work

Although the ICAR is a specific tool for assessing collaboration ability in medical residents, future studies should investigate if the ICAR can be specifically tailored for individual Royal College medical specialities. One such study is underway at Memorial University in the orthopedic surgery department.

Futhermore, performing a similar study in a much larger tertiary care center – such as in Toronto, Montreal, Vancouver, etc. – could significantly increase the number of raters and ratings necessary to provide in depth statistical analyses.

Finally, many residents were interested in two-way assessments regarding collaboration where the residents would be able to assess the nurses and allied health professionals on their collaboration. A study involving two-way assessment and resident self-assessment would provide even further insight to the important, yet subjective non-medical expert CanMEDS roles. With two-way assessment a new dimension of team dynamics can also be explored and may provide valuable information on how to maximize team efficiency and improve interprofessional relationships.

References / Bibliography

Accreditation Council of Graduate Medical Education (ACGME) History – http://www.acgme.org/acgmeweb/tabid/116/About.aspx Accessed: Dec 1, 2012.

Alkharusi, H. (2012). Generalizability Theory: An Analysis of Variance Approach to Measurement Problems in Educational Assessment. *Journal of Studies in Education*. 2(1): 184-196

Amin Z, Eng, KH, Gwee M, Hoon, TC, and Rhoon KD. (2006). Addressing the needs and priorities of medical teachers through a collaborative intensive faculty development programme. *Medical Teacher* 28(1): 85-8

Arora, S., Sevdalis, N., Suliman, I., Athanasiou, T., Kneebone, R., & Darzi, A. (2009). What makes a competent surgeon? Experts' and trainees' perceptions of the roles of a surgeon. *American Journal of Surgery* 198(5): 726-32

Blaikie, N (200). *Analysing Quantitative Data*. London: Sage Publications

Brienza RS, Huot S, Holmboe E. (2004). Influence of Gender on the Evaluation of Internal Medicine Residents. *Journal of Women's Health* 13(1):77-83

Campbell JL, Roberts M, Wright C, Hill J, Greco M, Taylor M, Richards S. (2011) Factors associated with variability in the assessment of UK doctors' professionalism: analysis of survey results. *British Medical Journal*. 27;343:d6212

Canadian Residency Matching Service (CaRMS). (2010). Canadian Students Studying Medicine Abroad. Retrieved from URL: http://www.carms.ca/pdfs/2010_CSA_Report/CaRMS_2010_CSA_Report.pdf (Feb 14th, 2013)

Canadian Residency Matching Service (CaRMS). (2012). First Choice Discipline and Match Results of CMGs by Gender - Part 1 2012 R-1 Main Residency Match - First iteration. Retrieved from URL: http://www.carms.ca/pdfs/2012R1_MatchResults/13_Discipline%20Choice%20and%20Match%20Results%20of%20CDN%20Grads%20by%20Gender1_en.pdf (Feb 14th, 2013)

Canavan C, Holtman MC, Richmond M, and Katsufrakis PJ. (2010). The Quality of Written Comments on Professional Behaviors in a Developmental Multisource Feedback Program. *Academic Medicine* 85:S106-9

Chou S, Cole G, McLaughlin K, and LockyerJ. (2008) CanMEDS evaluation in Canadian postgraduate training programmes: tools used and programme director satisfaction. Med Ed 42: 879-886

Chou S, Lockyer J, Cole G, and McLaughlin K. (2009) Assessing post-graduate trainees in Canada: Are we achieving diversity in methods? Medical Teacher 31: e58-e63

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

Cohen L, Manion L, Morrison K. (2000). *Research Methods in Education*. 5th edition. London: RoutledgeFalmer

College of Family Physicians of Canada (CFPC) – www.cfpc.ca/principles Accessed: Dec 20th, 2012

Colliver JA, Vu, NV, Marcy ML, Travis TA, and Robbs RS. (1993). Effects of examinee gender, standardized-patient gender, and their interaction on standardized patients' ratings of examinees' interpersonal and communication skills. *Academic Medicine* 68(2):153-157

Cook, DA & Beckman, T J (2009). Does scale length matter? A comparison of nine- versus five-point rating scales for mini-CEX. *Advances in Health Sciences Education*. 14: 655–684

Cronbach LJ. (1951). Coefficient alpha and the internal structure of tests. Psychometrika. 16, 297-334.

Cronbach LJ, Gleser GC, Nanda H, and Rajaranam N. (1972). *The dependability of behavorial measurements: Theory of Generalizability for scores and profiles*. Wiley, New York.

Curran, VS., Hollet, A., Casimiro, L.M., McCarthy, P., Banfield, VS., and Hall, P. (2011). Development and Validation of the Interprofessional Collaborator Assessment Rubric (ICAR). Interprofessional Care, 25: 339–344

Davis EC, Risucci RA, Blair PG, and Sachdeva, AK. (2011) Women in surgery residency programs: evolving trends from a national perspective. *J Am Coll Surg* 212(3):320-6

Day SC, Norcini JJ, Shea, JA, and Benson JA Jr. (1989). Gender Differences in the Clinical Competence of Resident in Internal Medicine. *Journal of General Internal Medicine* 4:309-12

Dent, JA, & Harden, RM (2009). *A Practical Guide for Medical Teacher (3rd edition)*. Churchill Livingstone: Elsevier

Dorsey JK & Colliver JA. (1995). Effect of anonymous test grading on passing rates as related to gender and race. *Academic Medicine* 70(4):321-3

Enders CK. (2010). *Applied Missing Data Analysis*. Guildford Press: New York, NY.

Ferguson E, James D, and Madeley L. (2002). Factors associated with success in medical school: systematic review of the literature. *British Medical Journal* 324:952

Fleenor J., & Prince J.M. (1997). Using 360-degree Feedback in Organizations: An Annotated Bibliography. Center for Creative Leadership. North Carolina.

Garra, G, Wackett, A, and Thode, H. (2011). Synchronous collection of multisource feedback evaluations does not increase inter-rater reliability.*Academic Emergency Medicine*. 18(2):S65-70

Glover Takahashi S, M Dawn and D Richardson. The CanMEDS Toolkit for Teaching and Assessing the Collaborator Role. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2012.

Gold MS & Bentler PM. (2000). Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. *Structural Equation Modeling* 7:3, 319-355

Gray, J. (1996) Primer on resident evaluation. *Annals RCPSC* 29:91–94

Hammick M, Freeth D, Koppel I, Reeves S, Barr H. (2007). A best evidence systematic review of interprofessional education: BEME Guide no. 9. Med Teach 29:735–751.

Hayes JR & Hatch JA (1999). Issues in measuring reliability. *Written Communication*.16(3):354-367.

Health Canada and Romanow, J. (2002). Building on Values: The Future of Health Care in Canada. Retrieved from: http://www.collectionscanada.gc.ca/webarchives/20071122004429/http://www.hc-sc.gc.ca/english/pdf/romanow/pdfs/hcc_final_report.pdf

Hill JJ, Ansprey A, Richards SH, and Campbell JL. (2012). Multisource feedback questionnaires in appraisal and for revalidation: a qualitative study in UK general practice. *Br J Gen Pract* 62(598):e314-21

Isaac C, Chertoff J, Lee B, Carne M. (2011). Do students' and authors' genders affect evaluations? A linguistic analysis of medical student performance evaluations. *Academic Medicine* 86(1):59-66

Jamieson, S (2004). Likert Scales: How to (Ab)Use Them. *Medical Education* 38: 1217-18.

Johnson D & Cujec, B (1989). Comparison of Self, Nurse, and Physician Assessment of Residents Rotating Through an Intensive Care Unit. Critical Care Medicine (11):1811-6.

Joshi R, Ling F, and Jaeger J. (2004). Assessment of a 360-Degree Instrument to Evaluate Residents' Competency in Interpersonal and Communication Skills. *Academic Medicine* 79(5): 458-63

Kaplan CB, Centor RM. (1990). The use of nurses to evaluate houseofficers' humanistic behavior. *Journal of General Internal Medicine* 5(5):410-4

Landis JR, Koch GG. (1977). The measurement of observer agreement for categorical data. *Biometrics.* 33:159-74.

LeBlanc, VR. (2009). The Effects of Acute Stress on Performance: Implications for Health Professions Education. *Academic Medicine* 84(10): S25-33

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology* 140:1–55

Lockyer J. (2003). Multisource Feedback in the Assessment of Physician Competencies. *J Contin Educ Health Prof* 23(1):4-12.

Lockyer JM, Violato C, Fidler H, and Alakija P. (2009). The Assessment of Pathologists/Laboratory Medicine Physicians Through a Multisource Feedback Tool. *Arch Pathol Lab Med* 133:1301-8

McMillan, JH & Schumacher S. (2006). *Research in Education: Evidence-Based Inquiry. 6th edition*. Pearson Education Inc, USA.

Manser, T. (2009). Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. Acta Anaesthesiol Scand. 53(2):143-51

Massagli TL & Carline JD (2007). Reliability of a 360-degree Evaluation to Assess Resident Competence. Am J Phys Med Rehabil 86(10):845-52

Moss-Racusin CA, Dovido, JF, Brescoll VL, Graham MJ, and Handelsman J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences.* 109(41): 16474-79

Musick DW, McDowell SM, Clark N, Salcido R. (2003). Pilot study of a 360-degree assessment instrument for physical medicine & rehabilitation residency programs. Am J Phys Med Rehabil. 82:394–402.

Newfoundland and Labrador Governmental Dept. of Education. Assessing and Evaluating Student Learning. Retrieved on December 6th, 2012, URL: http://www.ed.govs.nl.ca/edu/k12/curriculum/guides/english/primary/studentaccess.pdf

Nowack, KM (2011). Why 360 Degree Feedback Doesn't Work and What to Do About It. Presented at the International Personnel Assessment Conference (IPAC). Retrieved online on Jan 25, 2012. URL: http://www.ipacweb.org/conf/11/nowack.pdf

Ogunyemi D, Gonzalez G, Fong A, Alexander C, Finke D, Donnon T, Azziz R. (2009). From the eye of the nurses:360-degree evaluation of residents. *J ContinEduc Health Prof.* 29(2):105-10.

Ostroff C, Atwater LE, and Feinberg BJ. (2004). Understanding self-other agreement: A look at rater and ratee characteristics, Context, and outcomes. *Personnel Psychology* 57: 333-75

Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KM, Wollersheim HC, Grol R. (2007). Doctor performance assessment in daily practice: does it help doctors or not? A systematic review. *Medical Education* 41(11):1039–49.

Overeem K, Wollersheim H, Driessen E, Lombarts K, van de Ven G, Grol R, Arah O. (2009) Doctors' perceptions of why 360-degree feedback does (not) work: a qualitative study. *Medical Education* 43(9):874-82.

Overeem K, Wollersheim H, Onyebuchi AA, CruisbergJK, Grol RPTM, and Lombarts K. (2012). Evaluation of physicians' professional performance: An iterative development and validation study of multisource feedback instruments. *BMC Health Services Research* 12(80).

Parfrey, P. & Barrett, B (2009). *Clinical Epidemiology: Practice and Methods*. Humana Press, New York.

Porter SR, Whitcomb ME, and Weitzer WH. (2004). Multiple Surveys of Students and Survey Fatigue. *New Directions for Institutional Research* 121:63-73

Rand VE, Hudes ES, Browner WS, Wachter RM, and Avins AL. (1998). Effect of Evaluator and Resident Gender on the American Board of Internal Medicine Evaluation Scores. *Journal of General Internal Medicine* 13:670-4

Rezler AG, Bruce NC, Schmitt BP (1986). Dilemmas in the evaluation of Residents. Proceedings of the Annual Conference on Resident Medical Education 25:371-378

Ringsted, C., Hansen, T.L., & Scherpbier, A. (2006). Are some of the challenging aspects of the CanMEDS roles valid outside Canada? *Medical Education* 40:807-815

Royal College of Physicians and Surgeons of Canada (2005). CanMEDS Framework.. Retrieved from: http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/framework/the_7_canmeds_roles_e.pdf Accessed on: Jan 9, 2012.

Royal College of Physicians and Surgeons of Canada. URL: www.royalcollege.ca Accessed on: Dec 10th, 2012

Sargeant J, Mann K, & Ferrier S (2005). Exploring family physicians' reactions to multi-source feedback: perceptions of credibility and usefulness. *Medical Education* 39(5):497–504.

Sax LJ, Gilmartin SK, & Bryant AN. (2003). Assessing Response Rates and Nonresponse Bias in Web and Paper Surveys. *Research in Higher Education* 44(4):409-32

Stark R, Korenstein D, & Karani R (2008). Impact of a 360-degree Professionalism Assessment on Faculty Comfort and Skills in Feedback Delivery. *Journal of General Internal Medicine* 23(7):969-72

Stutsky BJ, Singer M, Renaud R (2012). Determining the weighting and relative importance of CanMEDS roles and competencies. BioMed Central Research Notes 5:354

Smith CJ, Rodenhuaser P, and Markert RJ. (1991). Gender bias of Ohio physicians in the evaluation of the personal statements of residency applicants. *Academic Medicine* 66(8):479-81

Smither & Walker (2004). Are the Characteristics of Narrative Comments Related to Improvement in Multirater Feedback Ratings Over Time? Personnel Psychology, 89, 575-581

Streiner DL & Norman GF (1989). *Health Measurement Scales: A Practical Guide to Their Development and Use.* Oxford University Press: New York.

Tan, S.C., Marlow, N., Field, J., Altree, M., Babidge, W., Hewett, P., and Maddern, G.J. (2012) A randomized crossover trial examining low- versus high-fidelity simulation in basic laparoscopic skills training. *Surgical Endoscopy* 26(11):3207-14

Thackeray EW, Halvorsen AJ, Ficalora RD, Engstler GJ, McDonald FS, and Oxentenko AS. (2012). The effects of gender and age on evaluation of trainees and faculty in gastroenterology. *American Journal of Gastroenterology* 107:1610-4

Trix F & Psenka C. (2003). Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. *Discourse and Society*. 14(2):191-220

Turnbull, J., Gray, J., and MacFadyen, J. (1998). Improving In-Training Evaluation Programs. *Journal of General Internal Medicine* 13(5): 317-323

Verma, S, Paterson, M, and Medves, J.(2006). Core Competencies for Health Care Professionals: What Medicine, Nursing, Occupational Therapy, and Physiotherapy Share. *Journal of Allied Health*. 35(2):109-115

Viera, AJ, and Garrett, JM. (2005). Understanding Intraobserver Agreement: The Kappa Statistic. *Family Medicine*. 37(5): 360-3

Violato C, Lockyer JM, and Fidler H. (2008). Assessment of Psychiatrists in Practice Through Multisource Feedback. *Can J Psychiatry* 53(8):525–33

Vivar, CG. (2006). Putting conflict management into practice: a nursing case study. *Journal of Nursing Management*. 14: 201-6

Warm EJ, Schauer D, Revis, B, and Boex JR. (2010). Multisource Feedback in the Ambulatory Setting. *Journal of Graduate Medical Education* 2(2):269-77

Webb, T.P. and Merkley, T.R. (2012). An evaluation of the success of a surgical resident learning portfolio *Journal of Surgical Education* 69(1):1-7

Weigelt, JA, Brasel, KJ, Bragg, D, and Simpson D. (2004). The 360-degree evaluation: increased work for little return? *Curr Surg*. 61(6): 627-8

Whitehead, C.R., Austin, Z., and Hodges, B.D. (2011). Flower power: the armoured expert in the CanMEDS competency framework? Advances in Health Sciences Education: Theory and Practice 16(5):681-94

Wiskin CMD, Allan TF, and Skelton JR. (2004). Gender as a variable in the assessment of final year degree-level communication skills. *Medical Education* 38:129-137

Wood J, Collins J, Burnside ES, et al. (2004). Patient, faculty, and selfassessment of radiology resident performance: A 360-degree method of measuring professionalism and Interpersonal/communication skills. Academic Radiology 11:931–9.

Wood L, Hassell A, Whitehouse A, Bullock A, Wall D. (2006). A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. Medical Teacher 28:e185–e191.

World Health Organization (2006). *Working Together for Health*. Retrieved from http://www.who.int/whr/2006/whr06_en.pdf

World Health Organization Study Group on Interprofessional Education and Collaborative Practice. (2010). *Framework for Action on Interprofessional Education & Collaborative Practice Geneva*: World Health Organization. Retrieved from http://www.who.int/hrh/resources/framework_action/en/index.html

Appendices

Appendix A – Memorial University Faculty of Medicine ITER

Memorial University PGME Elective or Selective Evaluation	Evaluated by: Evaluating: Dates:					
COLLABORATOR						
	Not Applicable	Rarely Meets Reasonable Expectations	Inconsistently Meets Reasonable Expectations	Generally Meets Reasonable Expectations	Sometimes Exceeds Reasonable Expectations	Consistently Exceeds Reasonable Expectations
1. Recognizes and acknowledges roles and expertise of team members	☐	☐	☐	☐	☐	☐
2. Interacts effectively with other team members	☐	☐	☐	☐	☐	☐

Appendix B – ICAR (Original Format)

Interprofessional Collaborator Assessment Rubric

Vernon Curran, MEd, PhD, Memorial University
Lynn Casimiro, PhD, PT, Montfort Hospital
Valerie Banfield, RN, MN, Registered Nurses Professional Development Centre
Pippa Hall, MD, CCFP, MEd, FCFP, University of Ottawa
Tracy Gierman, MA, Academic Health Council-Champlain Region
Kelly Lackie, RN, MN, CNCC(C), Registered Nurses Professional Development Centre
Ivy Oandasan, MD, MHSc, CCFP, FCFP, University of Toronto
Brian Simmons, BM, FRCPC, University of Toronto
Susan Wagner, MSc(CD), Reg. SLP(C), University of Toronto

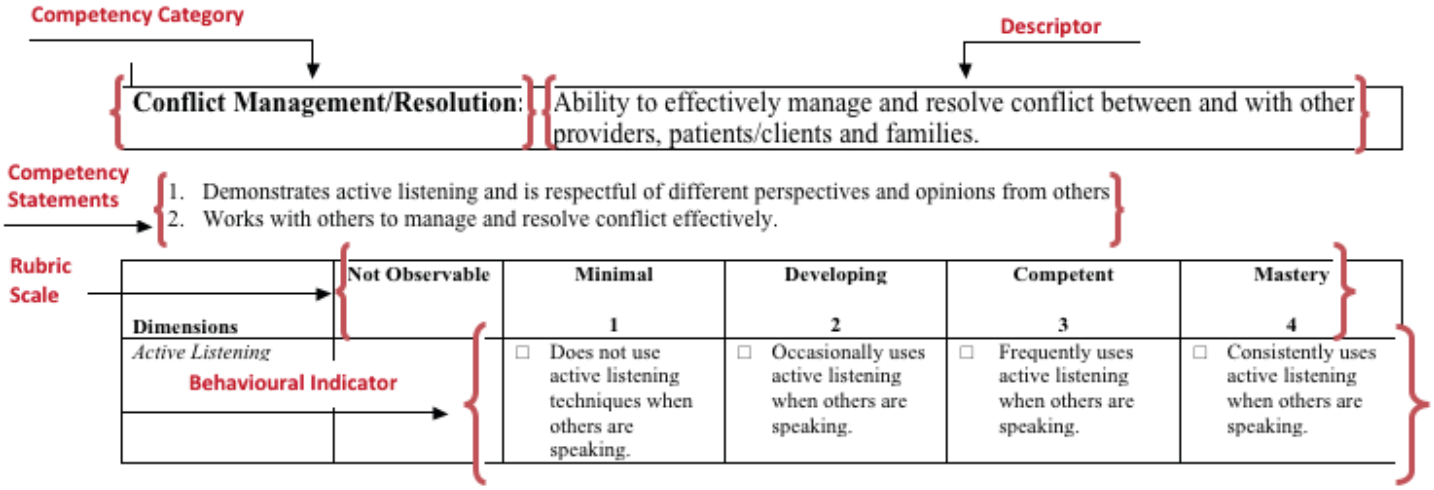
Project fund

What is a Rubric?

A Rubric is an assessment tool that lists a set of performance criteria which define and describe the important competencies being assessed. Rubrics are useful to instructors because it can improve the planning of learning experiences and increase the quality of direct instruction by providing focus, emphasis, and attention to particular details as a model for learners.

For learners, a rubric provides clear targets of proficiency to aim for. Learners can use Rubrics for self-assessment as individuals, in groups, and for peer assessment. It is believed that Rubrics may improve learners' performance and therefore increase learning, particularly when learners receive Rubrics beforehand, understand how they will be evaluated and can prepare accordingly. Rubrics are becoming increasingly popular with educators moving toward more authentic, performance-based assessments.

Interprofessional Collaborator Assessment Rubric



Using the Collaborator Rubric

The Interprofessional Collaborator Assessment Rubric is intended for use in the assessment of interprofessional collaborator competencies. Collaborative practice in health care occurs when multiple health workers from different professional backgrounds provide comprehensive services by working with patients, their families, carers and communities to deliver the highest quality of care across settings (WHO, 2010)[1]. Development of the Rubric tool was guided by an interprofessional advisory committee comprising educators from the fields of medicine, nursing and the rehabilitative sciences.

Key Principles

1) The Rubric has been developed for usage across different health professional education programs and in different learning contexts.

2) The Rubric dimensions are not intended to coincide with a specific year or level of a learner in his/her program of studies.

3) The Rubric may be used as a tool for formative and summative assessment of learners' competencies in interprofessional collaboration. As a formative assessment, the Rubric would allow learners to receive constructive feedback on competency areas for further development and improvement. As a summative assessment, the Rubric may be used to assess learners' achievement. The Rubric may also be introduced early in a program and used repeatedly to assess growth and development over time.

4) Usage of the Rubric in a reliable manner may require multiple interactions and repeated observation of a learner over a period of time.

5) Programs/disciplines should define remediation opportunities for learners not achieving an acceptable level of competency within their program area.

Rubric Validity

The Rubric dimensions are based on interprofessional collaborator competency statements that were developed and validated through a typological analysis of national and international competency frameworks, a Delphi survey of experts, and interprofessional focus groups with students and faculty.

[1] World Health Organization (WHO) Study Group on Interprofessional Education and Collaborative Practice. (2010). *Framework for Action on Interprofessional Education & Collaborative Practice.* Geneva, Switzerland: World Health Organization.

Interprofessional Collaborator Assessment Rubric

Instructions: For each of the dimensions below, check specific phrases which describe the performance of the learner.

Notes:
Assess by what is appropriate to the context/task.
- Occasionally: the learner demonstrates the desired behaviour once in a while.
- Frequently: the learner demonstrates the desired behaviour most of the time.
- Consistently: the learner always demonstrates the desired behaviour.

Communication: Ability to communicate effectively in a respectful and responsive manner with others ("others" includes team members, patient/client, and health providers outside the team).

1. Communicates and expresses ideas in an assertive and respectful manner.
2. Uses communication strategies (e.g. oral, written, information technology) in an effective manner with others.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Respectful Communication		☐ Communicates with others in a disrespectful manner.	☐ Occasionally communicates with others in a confident, assertive and respectful manner.	☐ Frequently communicates with others in a confident, assertive and respectful manner.	☐ Consistently communicates with others in a confident, assertive and respectful manner.
		☐ Does not communicate opinion or pertinent views on patient care with others.	☐ Occasionally communicates opinion or pertinent views on patient care with others.	☐ Frequently communicates opinion and pertinent views on patient care with others.	☐ Consistently communicates opinion and pertinent views on patient care with others.
		☐ Does not respond or reply to requests.	☐ Occasionally responds or replies to requests in a timely manner.	☐ Frequently responds or replies to requests in a timely manner.	☐ Consistently responds or replies to requests in a timely manner.
Communication Strategies		☐ Does not use communication strategies (verbal & non-verbal) appropriately with others.	☐ Occasionally uses communication strategies (verbal & non-verbal) appropriately.	☐ Frequently uses communication strategies (verbal & non-verbal) appropriately in a variety of situations.	☐ Consistently uses communication strategies (verbal & non-verbal) appropriately in a variety of situations.
		☐ Communication is illogical and unstructured.	☐ Occasionally communicates in a logical and structured manner.	☐ Frequently communicates in a logical and structured manner.	☐ Consistently communicates in a logical and structured manner.
		☐ Does not explain discipline-specific terminology/jargon.	☐ Occasionally explains discipline-specific terminology/jargon.	☐ Frequently explains discipline-specific terminology/jargon.	☐ Consistently explains discipline-specific terminology/jargon.
		☐ Does not use strategies that are appropriate for communicating with individuals with impairments (e.g., hearing, cognitive).	☐ Occasionally uses strategies that are appropriate for communicating with individuals with impairments (e.g., hearing, cognitive).	☐ Frequently uses strategies that are appropriate for communicating with individuals with impairments (e.g., hearing, cognitive).	☐ Consistently uses strategies that are appropriate for communicating with individuals with impairments (e.g., hearing, cognitive).
Comments:					

Collaboration: Ability to establish/maintain collaborative working relationships with other providers, patients/clients and families.

1. Establishes collaborative relationships with others in planning and providing patient/client care.
2. Promotes the integration of information from others in planning and providing care for patients/clients.
3. Upon approval of the patient/client or designated decision-maker, ensures that appropriate information is shared with other providers.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Collaborative Relationship		☐ Does not establish collaborative relationships with others.	☐ Occasionally establishes collaborative relationships with others.	☐ Frequently establishes collaborative relationships with others.	☐ Consistently establishes collaborative relationships with others.
Integration of Information from others		☐ Does not integrate information from others in planning and providing patient/client care.	☐ Occasionally integrates information from others in planning and providing patient/client care.	☐ Frequently integrates information and perspectives from others in planning and providing patient/client care.	☐ Consistently integrates information and perspectives from others in planning and providing patient/client care.
Information Sharing		☐ Does not share information with other providers.	☐ Occasionally shares information with other providers that is useful for the delivery of patient/client care.	☐ Frequently shares information with other providers that is useful for the delivery of patient/client care.	☐ Consistently shares information with other providers that is useful for the delivery of patient/client care.
		☐ Does not seek approval of patient/client or designated decision-maker when information is shared.	☐ Occasionally seeks approval of the patient/client or designated decision-maker when information is shared.	☐ Frequently seeks approval of the patient/client or designated decision-maker when information is shared.	☐ Consistently seeks approval of the patient/client or designated decision-maker when information is shared.
Comments:					

Roles and Responsibility: Ability to explain one's own roles and responsibilities related to patient/client and family care (e.g. scope of practice, legal and ethical responsibilities); and to demonstrate an understanding of the roles, responsibilities and relationships of others within the team.

1. Describes one's own roles and responsibilities in a clear manner.
2. Integrates the roles and responsibilities of others with one's own to optimize patient/client care.
3. Accepts accountability for one's contributions.
4. Shares evidence-based and/or best practice discipline-specific knowledge with others.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Roles and Responsibilities		☐ Does not describe one's own role and responsibilities with the team/patient/family.	☐ Occasionally describes one's own role and responsibilities with the team/patient/family.	☐ Frequently describes one's own roles and responsibilities with the team/patient/family.	☐ Consistently describes one's own roles and responsibilities in a clear manner with the team/patient/family.
Role/Responsibility Integration		☐ Does not include the roles and responsibilities of other providers in the delivery of patient care.	☐ Occasionally includes the roles and responsibilities of other providers in the delivery of patient care.	☐ Frequently includes the roles and responsibilities of all necessary health providers to optimize collaborative patient/client care.	☐ Consistently promotes and includes the roles and responsibilities of all necessary health providers to optimize collaborative patient/client care.
Accountability		☐ Does not demonstrate professional judgment when assuming tasks or delegating tasks.	☐ Occasionally demonstrates professional judgment when assuming tasks or delegating tasks.	☐ Frequently demonstrates professional judgment when assuming tasks or delegating tasks.	☐ Consistently demonstrates professional judgment when assuming tasks or delegating tasks.
		☐ Does not accept responsibility for the failure of collaborative goals.	☐ Occasionally accepts responsibility for the failure of collaborative goals.	☐ Frequently accepts responsibility for the failure of collaborative goals.	☐ Consistently accepts responsibility for the failure of collaborative goals.
		☐ Does not accept responsibility for individual actions that impact the team.	☐ Occasionally accepts responsibility for individual actions that impact the team.	☐ Frequently accepts responsibility for individual actions that impact the team.	☐ Consistently accepts responsibility for individual actions that impact the team.
		☐ Does not explain own scope of practice, code of ethics, standards and/or clinical guidelines in relation to collaborative patient-centred relationship.	☐ Occasionally explains own scope of practice, code of ethics, standards and/or clinical guidelines in relation to collaborative patient-centred relationship.	☐ Frequently explains own scope of practice, code of ethics, standards and/or clinical guidelines in relation to collaborative patient-centred relationship.	☐ Consistently explains own scope of practice, code of ethics, standards and/or clinical guidelines in relation to collaborative patient-centred relationship.
Sharing Evidence-Based/ Best Practice Knowledge		☐ Does not share evidence-based or best practice discipline-specific knowledge with others.	☐ Occasionally shares evidence-based or best practice discipline-specific knowledge with others.	☐ Frequently shares evidence-based or best practice discipline-specific knowledge with others.	☐ Consistently shares evidence-based or best practice discipline-specific knowledge with others.
Comments:					

Collaborative Patient/Client-Family Centred Approach: Ability to apply patient/client-centred principles through interprofessional collaboration.

1. Seeks input from patient/client and family in a respectful manner regarding feelings, beliefs, needs and care goals.
2. Integrates patient's/client's and family's life circumstances, cultural preferences, values, expressed needs, and health beliefs/behaviours into care plans.
3. Shares options and health care information with patients/clients and families.
4. Advocates for patient/client and family as partners in decision-making processes.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Patient/Client Input		☐ Does not seek input from patient/client and family.	☐ Occasionally seeks input from patient/client and family.	☐ Frequently seeks input from patient/client and family.	☐ Consistently seeks input from patient/client and family.
Integration of Patient/Client Beliefs and Values		☐ Does not integrate patient's/client's and family's circumstances, beliefs and values into care plans.	☐ Occasionally integrates the patient's/client's and family's circumstances, beliefs and values into care plans.	☐ Frequently integrates patient's/client's and family's circumstances, beliefs and values into care plans.	☐ Consistently promotes and integrates patient's/client's and family's circumstances, beliefs and values into care plans.
Information Sharing with Patient/Client		☐ Does not share options and health care information with patients/clients and families.	☐ Occasionally shares options and health care information with patients/clients and families.	☐ Frequently shares options and health care information with patients/clients and families.	☐ Consistently shares options and health care information with patients/clients and families.
Patient Advocacy in Decision-Making		☐ Does not advocate for patient/client and family as partners in decision-making processes.	☐ Occasionally advocates for patient/client and family as partners in decision-making processes.	☐ Frequently advocates for patient/client and family as partners in decision-making processes.	☐ Consistently advocates for patient/client and family as partners in decision-making processes.
Comments:					

Team Functioning: Ability to contribute to effective team functioning to improve collaboration and quality of care.

1. Recognizes and contributes to effective team functioning and dynamics.
2. Recognizes that leadership within the healthcare team may alternate or be shared depending on the situation.
3. Contributes in interprofessional team discussions.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Team Functioning and Dynamics		☐ Does not recognize the relationship between team functioning and quality of care.	☐ Occasionally demonstrates recognition of the relationship between team functioning and quality of care.	☐ Frequently demonstrates recognition of the relationship between team functioning and quality of care.	☐ Consistently demonstrates recognition of the relationship between team functioning and quality of care.
		☐ Does not recognize strategies that will improve team functioning.	☐ Occasionally demonstrates recognition of strategies that will improve team functioning.	☐ Frequently demonstrates recognition of strategies that will improve team functioning.	☐ Consistently demonstrates recognition of strategies that will improve team functioning.
Shared Leadership		☐ Does not recognize the importance of alternating or sharing leadership with others.	☐ Occasionally shares leadership and alternates leadership with others when appropriate for the discipline involved.	☐ Frequently shares leadership and alternates leadership with others when appropriate for the discipline involved.	☐ Consistently shares leadership and alternates leadership with others when appropriate for the discipline involved.
Team Discussion		☐ Does not view themselves as part of the team.	☐ Occasionally demonstrates recognition of themselves as part of a team.	☐ Frequently demonstrates recognition of themselves as part of a team.	☐ Consistently demonstrates recognition of themselves as part of a team.
		☐ Does not contribute to interprofessional team discussions.	☐ Occasionally contributes to interprofessional team discussions.	☐ Frequently contributes to interprofessional team discussions.	☐ Consistently contributes to interprofessional team discussions.
Comments:					

Conflict Management/Resolution: Ability to effectively manage and resolve conflict between and with other providers, patients/clients and families.

1. Demonstrates active listening and is respectful of different perspectives and opinions from others.
2. Works with others to manage and resolve conflict effectively.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Respect for different perspectives		☐ Does not consider the perspectives and opinions of others.	☐ Occasionally seeks the perspectives and opinions of others.	☐ Frequently seeks the perspectives and opinions of others.	☐ Consistently seeks the perspectives and opinions of others.
		☐ Does not seek clarification in a respectful manner when misunderstandings arise.	☐ Occasionally seeks clarification when misunderstandings arise, but it is not necessarily done in a respectful manner.	☐ Frequently seeks clarification in a respectful manner when misunderstandings arise.	☐ Consistently seeks clarification in a respectful manner when misunderstandings arise.
Active Listening		☐ Does not use active listening techniques when others are speaking.	☐ Occasionally uses active listening when others are speaking.	☐ Frequently uses active listening when others are speaking.	☐ Consistently uses active listening when others are speaking.
Conflict Management		☐ Does not manage or resolve conflict with others.	☐ Occasionally uses appropriate conflict resolution strategies to manage and/or resolve conflict.	☐ Frequently uses appropriate conflict resolution strategies to manage and/or resolve conflict.	☐ Consistently uses appropriate conflict resolution strategies to manage and/or resolve conflict.
Comments:					

Appendix C – ICAR (Modified 4-point scale)

Interprofessional Collaborator Assessment Rubric

for Anaesthesiology Residency

Adapted from:

Curran, V.R., Casimiro, L., Banfield, V., Hall, P., Lackie, K., Simmons, B., Tremblay, M., Wagner, S.J., Oandasan, I. (2011). *Development and Validation of the Interprofessional Collaborator Assessment Rubric (ICAR).* Journal of Interprofessional Care, 25, 339-344.

Interprofessional Collaborator Assessment Rubric

Instructions: For each of the dimensions below, check specific phrases which describe the performance of the learner.

Notes:

Assess by what is appropriate to the context/task.

- Occasionally: the learner demonstrates the desired behaviour once in a while.
- Frequently: the learner demonstrates the desired behaviour most of the time.
- Consistently: the learner always demonstrates the desired behaviour.

Communication: Ability to communicate effectively in a respectful and responsive manner with others ("others" includes team members, patient/client, and health providers outside the team).

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Respectful Communication		☐ Communicates with others in a disrespectful manner.	☐ Occasionally communicates with others in a confident, assertive and respectful manner.	☐ Frequently communicates with others in a confident, assertive and respectful manner.	☐ Consistently communicates with others in a confident, assertive and respectful manner.
		☐ Does not communicate opinion or pertinent views on patient care with others.	☐ Occasionally communicates opinion or pertinent views on patient care with others.	☐ Frequently communicates opinion and pertinent views on patient care with others.	☐ Consistently communicates opinion and pertinent views on patient care with others.
Communication Strategies		☐ Does not use communication strategies (verbal & non-verbal) appropriately with others.	☐ Occasionally uses communication strategies (verbal & non-verbal) appropriately.	☐ Frequently uses communication strategies (verbal & non-verbal) appropriately in a variety of situations.	☐ Consistently uses communication strategies (verbal & non-verbal) appropriately in a variety of situations.
		☐ Communication is illogical and unstructured.	☐ Occasionally communicates in a logical and structured manner.	☐ Frequently communicates in a logical and structured manner.	☐ Consistently communicates in a logical and structured manner.

Collaboration: Ability to establish/maintain collaborative working relationships with other providers, patients/clients and families.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Collaborative Relationship		☐ Does not establish collaborative relationships with others.	☐ Occasionally establishes collaborative relationships with others.	☐ Frequently establishes collaborative relationships with others.	☐ Consistently establishes collaborative relationships with others.
Integration of Information from others		☐ Does not integrate information from others in planning and providing patient/client care.	☐ Occasionally integrates information from others in planning and providing patient/client care.	☐ Frequently integrates information and perspectives from others in planning and providing patient/client care.	☐ Consistently integrates information and perspectives from others in planning and providing patient/client care.
Information Sharing		☐ Does not share information with other providers.	☐ Occasionally shares information with other providers that is useful for the delivery of patient/ client care.	☐ Frequently shares information with other providers that is useful for the delivery of patient/ client care.	☐ Consistently shares information with other providers that is useful for the delivery of patient/ client care.

Roles and Responsibility: Ability to explain one's own roles and responsibilities related to patient/ client and family care (e.g. scope of practice, legal and ethical responsibilities); and to demonstrate an understanding of the roles, responsibilities and relationships of others within the team.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Roles and Responsibilities		☐ Does not describe one's own role and responsibilities with the team/patient/ family.	☐ Occasionally describes one's own role and responsibilities with the team/patient/ family.	☐ Frequently describes one's own roles and responsibilities with the team/patient/ family.	☐ Consistently describes one's own roles and responsibilities in a clear manner with the team/patient/ family.
Accountability		☐ Does not demonstrate professional judgment when assuming tasks or delegating tasks.	☐ Occasionally demonstrates professional judgment when assuming tasks or delegating tasks.	☐ Frequently demonstrates professional judgment when assuming tasks or delegating tasks.	☐ Consistently demonstrates professional judgment when assuming tasks or delegating tasks.
Sharing Evidence-Based/ Best Practice Knowledge		☐ Does not share evidence-based or best practice discipline-specific knowledge with others.	☐ Occasionally shares evidence-based or best practice discipline-specific knowledge with others.	☐ Frequently shares evidence-based or best practice discipline-specific knowledge with others.	☐ Consistently shares evidence-based or best practice discipline-specific knowledge with others.

Collaborative Patient/Client-Family Centred Approach: Ability to apply patient/client-centred principles through interprofessional collaboration.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Patient/Client Input		☐ Does not seek input from patient/client and family.	☐ Occasionally seeks input from patient/client and family.	☐ Frequently seeks input from patient/client and family.	☐ Consistently seeks input from patient/client and family.
Information Sharing with Patient/Client		☐ Does not share options and health care information with patients/clients and families.	☐ Occasionally shares options and health care information with patients/clients and families.	☐ Frequently shares options and health care information with patients/clients and families.	☐ Consistently shares options and health care information with patients/clients and families.

Team Functioning: Ability to contribute to effective team functioning to improve collaboration and quality of care.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Team Functioning and Dynamics		☐ Does not recognize the relationship between team functioning and quality of care.	☐ Occasionally demonstrates recognition of the relationship between team functioning and quality of care.	☐ Frequently demonstrates recognition of the relationship between team functioning and quality of care.	☐ Consistently demonstrates recognition of the relationship between team functioning and quality of care.
Team Discussion		☐ Does not contribute to interprofessional team discussions.	☐ Occasionally contributes to interprofessional team discussions.	☐ Frequently contributes to interprofessional team discussions.	☐ Consistently contributes to interprofessional team discussions.

Conflict Management/Resolution: Ability to effectively manage and resolve conflict between and with other providers, patients/clients and families.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
Respect for different perspectives		☐ Does not consider the perspectives and opinions of others.	☐ Occasionally seeks the perspectives and opinions of others.	☐ Frequently seeks the perspectives and opinions of others.	☐ Consistently seeks the perspectives and opinions of others.
		☐ Does not seek clarification in a respectful manner when misunderstandings arise.	☐ Occasionally seeks clarification when misunderstandings arise, but it is not necessarily done in a respectful manner.	☐ Frequently seeks clarification in a respectful manner when misunderstandings arise.	☐ Consistently seeks clarification in a respectful manner when misunderstandings arise.
Conflict Management		☐ Does not manage or resolve conflict with others.	☐ Occasionally uses appropriate conflict resolution strategies to manage and/or resolve conflict.	☐ Frequently uses appropriate conflict resolution strategies to manage and/or resolve conflict.	☐ Consistently uses appropriate conflict resolution strategies to manage and/or resolve conflict.

Appendix D – ICAR (Modified 9-point scale)

MEMORIAL UNIVERSITY

Centre for Collaborative Health Professional Education

Interprofessional Collaborator Assessment Rubric

Thank you for participating in the Faculty of Medicine's research project focused on improving evaluation techniques of resident collaboration. Your participation will be completely anonymous.

This portion of our study focuses on the level of inter-rater agreement between medical professionals - physicians, nursing staff, and allied health professionals – using the Interprofessional Collaborator Assessment Rubric (ICAR).

We welcome all comments you may have regarding this assessment tool and ideas you may have regarding assessing interprofessional collaboration.

Thank you for your time and participation in this worthwhile research project.

Sincerely,

Mark Hayward B.Sc. B.Ed.
--
Clinical Epidemiology M.Sc. Candidate
in Medical Education at Memorial University

Please check the corresponding responses that pertain to you

Question #1 Select category indicating your profession:

☐ Physician ☐ RN ☐ LPN ☐ PT ☐ OT ☐ Social Work ☐ Pharmacy

☐ Speech Language Pathologist ☐ Dietician ☐ Other_____

Question #2 Gender: ☐ Male ☐ Female

Question #3 Total years of experience in your **current profession**

☐ Less than 1 ☐ 2 – 5 ☐ 6 – 10 ☐ 11 – 15 ☐ 16 – 20 ☐ 21+

Question #4 Total years of experience in this **medical / surgical unit**

☐ Less than 1 ☐ 2 – 5 ☐ 6 – 10 ☐ 11 – 15 ☐ 16 – 20 ☐ 21+

Question #5 Approximately how often did you interact with the resident being evaluated?

☐ Multiple times per shift ☐ Once per shift ☐ Several times per week ☐ Rarely

Question #6 Describe the types of interactions you had with the resident being evaluated (check all that apply)

☐ Direct (face to face)
☐ Phone consultation
☐ Via chart notes / orders / requests
☐ Discharge planning
☐ Hearing from other colleague's interactions with resident
☐ Hearing from patient or family member's interactions with resident

MEMORIAL UNIVERSITY

Centre for Collaborative Health Professional Education

Interprofessional Collaborator Assessment Rubric

Instructions: For each of the statements below, circle the number which corresponds to the performance of the learner.

1	2	3	4	5	6	7	8	9	N/O
Well Below Expected		Below Expected		Expected	Above Expected		Well Above Expected		Not Observable

Communication: Ability to communicate effectively in a respectful and responsive manner with others ("others" includes team members, patient/client, and health providers outside the team).

Resident...										N/O
Communicates with others in a confident, assertive, and respectful manner.	1	2	3	4	5	6	7	8	9	☐
Communicates opinion and pertinent views on patient care with others.	1	2	3	4	5	6	7	8	9	☐
Uses communication strategies (verbal & non-verbal) appropriately in a variety of situations.	1	2	3	4	5	6	7	8	9	☐
Communicates in a logical and structured manner	1	2	3	4	5	6	7	8	9	☐

Collaboration: Ability to establish/maintain collaborative working relationships with other providers, patients/clients and families.

Resident...										N/O
Establishes collaborative relationships with others.	1	2	3	4	5	6	7	8	9	☐
Integrates information and perspectives from others in planning and providing patient/client care.	1	2	3	4	5	6	7	8	9	☐
Shares information with other providers that is useful for the delivery of patient/client care.	1	2	3	4	5	6	7	8	9	☐

Roles and Responsibility: Ability to explain one's own roles and responsibilities related to patient/ client and family care (e.g. scope of practice, legal and ethical responsibilities); and to demonstrate an understanding of the roles, responsibilities and relationships of others within the team.

Resident...										N/O
Describes one's own roles and responsibilities in a clear manner with the team/patient/family.	1	2	3	4	5	6	7	8	9	☐
Demonstrates professional judgement when assuming or delegating tasks.	1	2	3	4	5	6	7	8	9	☐
Shares evidence-based or best practice discipline-specific knowledge with others.	1	2	3	4	5	6	7	8	9	☐

MEMORIAL UNIVERSITY

Centre for Collaborative Health Professional Education

Collaborative Patient/Client-Family Centred Approach: Ability to apply patient/client-centred principles through interprofessional collaboration.

Resident...										N/O
Seeks input from patient/client and family.	1	2	3	4	5	6	7	8	9	☐
Shares options and health care information with patients/clients and families.	1	2	3	4	5	6	7	8	9	☐

Team Functioning: Ability to contribute to effective team functioning to improve collaboration and quality of care.

Resident...										N/O
Demonstrates recognition of the relationship between team functioning and quality of care.	1	2	3	4	5	6	7	8	9	☐
Contributes to interprofessional team discussions.	1	2	3	4	5	6	7	8	9	☐

Conflict Management/Resolution: Ability to effectively manage and resolve conflict between and with other providers, patients/clients and families.

Resident...										N/O
Seeks the perspectives and opinions of others.	1	2	3	4	5	6	7	8	9	☐
Seeks clarification in a respectful manner when misunderstandings arise.	1	2	3	4	5	6	7	8	9	☐
Uses appropriate conflict resolution strategies to manage and/or resolve conflict.	1	2	3	4	5	6	7	8	9	☐

With **respect to collaboration ability**, compared to other residents you have previously interacted with, this resident was:

1	2	3	4	5	6	7	8	9
Well Below Average		Below Average		Average	Above Average		Well Above Average	

Comments regarding the resident's collaboration ability:_____

_____

_____

Comments regarding the study or ICAR:_____

_____

_____