UNDERSTANDING HOW RATERS COMMUNICATE IN THE CONTEXT OF MEDICAL HIGH-STAKES, PERFORMANCE-BASED ASSESSMENTS

By

Stefanie S. Sebok

Dissertation submitted to the Faculty of Education in conformity with the requirements for the degree of Doctor of Philosophy

Queen’s University
Kingston, Ontario, Canada
September, 2015

Copyright © Stefanie S. Sebok, 2015
ABSTRACT

It is well established in the medical education literature that raters represent a significant source of unexplained variance in performance assessments. To better understand the variance associated with raters’ evaluative judgments, this dissertation investigated how raters, in various assessment contexts, communicate and articulate what they observe during performance assessments. Empirical data collected from three different performance assessments: a multiple mini-interview (MMI), an objective structured clinical examination (OSCE), and workplace-based assessments (WBAs) were analyzed to better understand raters’ behaviours. Relationships that exist between different forms of assessment were also examined. The first study established a proof of concept, showing that raters are not always consciously aware of the associations they make during assessments. The second study examined conditions under which raters find it difficult to make a seemingly uncomplicated judgment. Finally, the third study examined relationships that exist between quantitative and qualitative methods used in the evaluation of performance assessments. The empirical findings from this dissertation suggest that although raters may behave idiosyncratically, there exist patterns of behaviour that could be used to attain better information from raters about a learner’s performance and potentially allow raters to be more suitably matched according to their strengths in performance-based assessments. The findings of this research should help to overcome some of the challenges associated with rater-mediated assessments.
DEDICATION

I dedicate this work to Dr. Peter D. MacMillan, the man who first recognized my talent, encouraged and supported me every step of the way, and never stopped being my biggest fan.
ACKNOWLEDGEMENTS

Don Klinger, my dissertation supervisor, you made me wiser by making me work a little bit harder and learn a little bit faster, thanks for making me that much stronger.

Sue Fostaty-Young and George Engelhard Jr., my dissertation committee members, thanks for all the time you invested in reading and providing feedback on the dissertation.

Tim Wood and Deborah Tregunno, my external examiners, thank you for posing challenging, thought provoking questions at the defense; and Amanda Cooper, the Dean’s Delegate, thanks for sharing your knowledge and expertise.

Thank you to the institutions and participants that took part in my studies and to the Social Sciences and Humanities Research Council for funding my research.

To all my colleagues at Medical Council of Canada, especially Sydney Smee, Marguerite Roy, and André De Champlain, who brought me into medical education and cheered me on.

To my mentors at Pearson Research and Innovation Network, Ed Wolfe, Katie McClarty, and Kimberly O’Malley, thank you for providing me with industry experience and exposing me to new ways of approaching educational challenges.

Andrea Gingerich, sharing in the Ph.D experience with you has certainly bonded us for life.

Thank you for being there whenever I needed you, sharing your medical education knowledge, and telling me to put on my “big-girl pants” and just get it done.

To my family and friends, especially my mom – Susane Sebok, who have always been there providing support, encouragement, and love.

Finally, to Mark D. Syer, one of the two men in this world who think I walk on water, thank you for being everything that I am not and sharing those qualities so that I can be a better person. I am so grateful to have you and I look forward to continuing to grow with you.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>ix</td>
</tr>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Research Purpose</td>
<td>4</td>
</tr>
<tr>
<td>Outlining the Learning Contexts for this Research</td>
<td>5</td>
</tr>
<tr>
<td>The Admissions Context</td>
<td>5</td>
</tr>
<tr>
<td>The Undergraduate Context</td>
<td>6</td>
</tr>
<tr>
<td>The Postgraduate Context</td>
<td>8</td>
</tr>
<tr>
<td>Dissertation Overview</td>
<td>10</td>
</tr>
<tr>
<td>CHAPTER 2: LITERATURE REVIEW AND THEORETICAL FRAMEWORK</td>
<td>14</td>
</tr>
<tr>
<td>Understanding the Term – “Rater Effects”</td>
<td>14</td>
</tr>
<tr>
<td>Raters are Error-Prone</td>
<td>17</td>
</tr>
<tr>
<td>Raters’ Cherished Values</td>
<td>17</td>
</tr>
<tr>
<td>Raters’ Emotional Responses</td>
<td>18</td>
</tr>
<tr>
<td>Raters’ Impression Formation</td>
<td>19</td>
</tr>
<tr>
<td>Raters’ Mental Workload Capacity</td>
<td>20</td>
</tr>
<tr>
<td>Dual Processing Memory Models</td>
<td>21</td>
</tr>
<tr>
<td>Raters can be Educated</td>
<td>22</td>
</tr>
<tr>
<td>Differences in Raters’ Frames of Reference</td>
<td>22</td>
</tr>
<tr>
<td>Differences in Raters’ Experiences</td>
<td>23</td>
</tr>
<tr>
<td>Differences in the Use of Scales and Instruments</td>
<td>24</td>
</tr>
<tr>
<td>Differences in Rater Training</td>
<td>25</td>
</tr>
<tr>
<td>Raters can Provide Different, but Relevant Pieces of Information</td>
<td>26</td>
</tr>
<tr>
<td>Different Ways of Understanding Rater Idiosyncrasy</td>
<td>26</td>
</tr>
<tr>
<td>Variation due to Different Perceptions of the Same Performance</td>
<td>27</td>
</tr>
<tr>
<td>Conclusion</td>
<td>28</td>
</tr>
<tr>
<td>Theoretical Framework</td>
<td>30</td>
</tr>
<tr>
<td>CHAPTER 3: SEEING THINGS DIFFERENTLY OR SEEING DIFFERENT THINGS?</td>
<td>35</td>
</tr>
<tr>
<td>EXPLORING RATERS’ ASSOCIATIONS OF NON-COGNITIVE ATTRIBUTES</td>
<td>35</td>
</tr>
<tr>
<td>Abstract</td>
<td>35</td>
</tr>
<tr>
<td>Seeing Things Differently or Seeing Different Things? Exploring Raters’ Associations of Non-Cognitive Attributes</td>
<td>36</td>
</tr>
<tr>
<td>Methods</td>
<td>38</td>
</tr>
<tr>
<td>Sample</td>
<td>38</td>
</tr>
<tr>
<td>Rater training</td>
<td>39</td>
</tr>
<tr>
<td>Instruments and Procedure</td>
<td>39</td>
</tr>
<tr>
<td>Rasch analyses</td>
<td>39</td>
</tr>
<tr>
<td>Hierarchical clustering</td>
<td>40</td>
</tr>
<tr>
<td>Results</td>
<td>41</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

## CHAPTER 3
- **Table 3.1**  
  Content and Attributes for MMI Stations  
  Page 49
- **Table 3.2**  
  Rasch Reliability Coefficients for Applicants, raters, and Attributes according to Station  
  Page 49

## CHAPTER 4
- **Table 4.1**  
  Descriptive Information and Reliability Indices for OSCE Stations  
  Page 59
- **Table 4.2**  
  Rater Differences between Binary and Three-Point Checklist Tools  
  Page 60
- **Table 4.3**  
  Descriptive Measures for Differences in the Scoring of “Attempted” among Candidates  
  Page 61

## CHAPTER 5
- **Table 5.1**  
  Regression Predictors for Checklist Option, “Done, but needs attention”  
  Page 79
# LIST OF FIGURES

## CHAPTER 2

**Figure 2.1**  
The Four Quadrants Derived from the Dimensionally Based Categorization Perspective  
31

## CHAPTER 3

**Figure 3.1**  
Dendrogram for the comparison of faculty and student raters and Station 4. The height at which two attributes are joined represents how dissimilar they are. For example, from the dendrogram on the left (Station 4 - Faculty), we find that Effectiveness and Resolution are the most similar (they are "joined" at the lowest height) while Professionalism is the most distinct attribute (it is "joined" at the greatest height).  
50

**Figure 3.2**  
Dendrogram for the comparison of faculty and student raters at Station 6  
50

**Figure 3.3**  
Dendrogram comparing the faculty and student rater that simultaneously observed and rated applicants at Station 4  
51

**Figure 3.4**  
Dendrogram comparing the faculty and student rater that simultaneously observed and rated applicants at Station 6  
51

## CHAPTER 4

**Figure 4.1**  
Overview of the Overall Research Design  
56
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>CIV</td>
<td>Construct Irrelevant Variance</td>
</tr>
<tr>
<td>EPA</td>
<td>Entrustable Professional Activity</td>
</tr>
<tr>
<td>McMAP</td>
<td>McMaster Modular Assessment Program</td>
</tr>
<tr>
<td>MFRM</td>
<td>Many-Facet Rasch Model</td>
</tr>
<tr>
<td>MMI</td>
<td>Multiple Mini-Interview</td>
</tr>
<tr>
<td>OSCE</td>
<td>Objective Structured Clinical Examination</td>
</tr>
<tr>
<td>SSHRC</td>
<td>Social Sciences and Humanities Research Council</td>
</tr>
<tr>
<td>WBA</td>
<td>Workplace-Based Assessment</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION

The shift towards competency-based education has created a movement that favours developing and ensuring that medical professionals are able to meet the needs of patients in a safe, efficient, and professional manner (Hodges & Lingard, 2012). Underlying the desire to produce competent medical professionals lies the challenge of being able to assess their medical competence. Assessment of medical competence primarily occurs in two ways: 1) through traditional written examinations, which are designed to evaluate clinical knowledge and reasoning, and 2) through performance-based assessments, which are intended to assess the practical skills of medical trainees or practitioners as they complete a series of simulated or context-specific medical scenarios. Within the paradigm of performance-based assessments, the multiple mini-interview (MMI), objective structured clinical examinations (OSCEs), and workplace-based assessments (WBAs) have come to dominate the field of medical education given their design to showcase practical applications (Linn & Burton, 1994).

Despite the appropriateness and support for performance-based assessments in the context of medical education, these assessments are resource-intensive to administer and evaluate. First and foremost these assessments require the clinical judgment of qualified medical professionals or other trained individuals (i.e., raters) who assess and measure the competence of medical trainees. These raters, usually observers of the simulated medical scenario, are required to make expert judgments about the quality of a candidate’s performance. However, within the context of performance assessments, raters are often identified as a main source of error and undesired variation (e.g., Harik et al., 2009; LaMantia et al., 1999; Sebok, Luu, & Klinger, 2014; cf. van der Vleuten & Swanson, 1990) because of their tendency to provide inconsistent or divergent information. These behaviours often threaten the validity (especially aspects related to
reliability and fairness) of results generated from high-stakes assessments (Sebok, Roy, Klinger, & De Champlain, 2015).

Performance assessments have been, and continue to be, an area of growing interest in medical education. When attempting to assess complex skills and general competencies, research has strongly supported the use of performance based-assessments (e.g., WBAs, OSCEs, and other simulation-based encounters). These assessments are believed to be more “authentic” in their design and structure, making them better able to measure clinical aspects (Linn & Burton, 1994) and assess complex knowledge and skills (Swanson, Norman, & Linn, 1995), especially when compared to traditional written examinations. However, despite the extensive amount of research and development on these forms of assessment, medical education has faced the realization that even performance assessments are not without challenges.

Although performance-based assessments have become an integral part of many assessment programs, they can be problematic because they rely on raters’ evaluative judgments. Any assessments that involve the use of raters are called rater-based assessments. Rater-based assessments, also sometimes referred to as rater-mediated assessments, receive much attention in the medical education literature. Medical education relies on rater-based assessments, especially in high-stakes assessment contexts, because the judgment and perspectives of current and practicing physicians are highly regarded. Medical and professional competence is usually measured against a specific standard or criterion; however, in the context of performance-based assessments, raters sometimes exhibit different interpretations of an individual’s performance in relation to the aforementioned criterion due to their own personal training and beliefs about what constitutes appropriate patient care (Kogan, Hess, Conforti, & Holmboe, 2010). These different
interpretations can lead to variability in judgments, and in some cases even incorrect decisions, about whether or not an individual is fit to practice medicine.

Research has previously identified both the potentials and challenges of rater-based assessments (e.g., Holmboe, Hawkins, & Hout, 2004; Sebok et al., 2014). Certainly, raters add to the element of subjectivity in performance assessments (Newble, 2004), which can be problematic because rater idiosyncrasy (i.e., raters’ individual distinctiveness) can impact both the reliability of an assessment and the validity of an assessment decision. To complicate this matter further, some researchers have taken the position that the resulting variability in raters’ judgments might actually be a good thing despite the impact it may have on the validity of assessments (see Gingerich, Regehr, & Eva, 2011). Given that medical trainees are primarily assessed through some form of direct observation of performance, it is imperative that further research be conducted to investigate not only variability in the quality of information obtained from raters, but also the extent to which their judgments may be compromising the precision and accuracy of performance-based assessments.

Construct irrelevant variance (CIV) describes any variability not related to the focus of measurement. CIV is caused by extraneous variables and has the potential to affect assessment outcomes; therefore, it is generally considered to represent error variance (Sebok, et al., 2015). In the case of judging performance assessments, understanding the sources of CIV requires a thorough comprehension and knowledge of how formulating a judgment leads to the development of an interpretation. While exploring how error variance is associated with rater variability, Crossley, Johnson, Booth, and Wade (2011) found that raters considered many assessment tools to be invalid in capturing the clinical complexities measured in performance assessments. Further investigation of the relationship between raters’ judgments and assessment
instruments revealed that assessments should be designed to reflect the cognitive processes of raters, encourage the subjective judgments of raters, and solicit the participation of raters who are most suited to evaluate clinical encounters (Crossley & Jolly, 2012). Notwithstanding, any one of these factors has the potential to introduce construct irrelevant variance in performance assessments.

In an attempt to increase the validity of performance-based assessments, several researchers have proposed the use of “programmatic assessment” (see Schuwirth & van der Vleuten, 2011; van der Vleuten et al., 2012). The premise underlying the use of programmatic assessment is that broadening the types of assessment measures and assessing individuals repeatedly over longer periods of time allows for a more accurate and fruitful representation of one’s medical competence. Therefore, monitoring and collecting assessment data from various learning contexts is imperative. One aspect that programmatic assessment fails to address is the fundamental issue associated with rater idiosyncrasy, which is that raters are not always able, or willing, to accurately identify and articulate their evaluative judgments. Without adequately addressing the problems associated with raters’ inability to translate their judgments, researchers and decision makers are often left trying to find a signal in the noise.

**Research Purpose**

This dissertation includes three studies that intended to address gaps in the research literature previously outlined. Specifically, the primary purpose of this research was to further our current understandings about how raters communicate and articulate evaluative judgments by examining their rating behaviour and identifying relationships that exist between various methods of assessment. This purpose was addressed across the three distinct, but related, empirical studies that focused on raters and their behaviours within the context of high-stakes
performance assessments. The first study, which took place in an admissions context, explored how raters make associations between various competencies and whether those associations contribute to rater variability. The second study, that took place in an undergraduate medical education context, examined how raters use alternative assessment formats to capture nuances in trainees’ behaviours. Finally, the third study, which took place in a postgraduate medical context, investigated whether allowing for multiple methods of assessment leads to better information from raters. This study also explored the alignment between various methods of assessment as understanding these relationships would be imperative to individuals implementing programmatic assessment practices.

Outlining the Learning Contexts for this Research

Programmatic assessment acknowledges that performance of individuals is largely context dependent (van der Vleuten et al., 2012). Therefore, in order to help make sense of the findings derived from the empirical studies, a brief overview of the three different study contexts (i.e., admissions, undergraduate, and postgraduate) is provided. Given that context has the ability to influence one’s interpretation of an event and lead to variability in assessments (Eva, 2003), it is important to consider the role of context when examining rater-based assessments.

The Admissions Context

Admissions procedures serve as a crucial form of gatekeeping. Given the resources and costs associated with professional programs such as medical school, once an applicant is admitted, institutions are highly motivated to keep the individual in the program (Dudek, Marks, & Regehr, 2005; Sebok & MacMillan, 2014) making the admissions process particularly important. Within the context of medical education, there exists a need to ensure that students entering medical school possess certain attributes and qualities that are necessary for them to be
successful not only in the program, but also in the medical profession. Hence, pre-entry admissions processes are critically important to identify those candidates who will most likely be successful in their subsequent professional program of study. Therefore, these admissions processes are high-stakes assessments not only due to the decision-making aspect, but also because of the long-term consequences of such decisions for those attempting to gain admission (Pau et al., 2013). Given that successful completion of medical school is one of the requirements needed to practice medicine, an individual’s inability to secure a seat in one of Canada’s 17 accredited medical programs ends, or at least delays, an applicant’s ability to become a licensed physician in Canada.

Medical institutions often have admissions processes that are highly competitive and thus high-stakes in nature. Within the last decade, the MMI has become an increasingly common admissions tool used at Canadian and international medical schools. The MMI, in an effort to measure traits such as critical thinking and empathy, requires medical applicants to participate in a series of mini-interviews that address specific dilemmas or social situations. Given that it is well-known that raters represent a main source of unexplained and often undesired variance in performance-based assessments, this study, conducted in the admissions context, was designed to better understand rater variance by investigating the relationships that exist between various non-cognitive attributes through an examination of how raters distinguish between different attributes in performance-based assessments.

The Undergraduate Context

Many undergraduate medical education programs, at least in Canada, use a cohort model for teaching and assessing medical trainees. A cohort model is when a certain number of medical trainees (usually around 100), all admitted in the same year, work their way through medical
school together taking the same courses according to a set schedule. Performance assessments used in undergraduate medical education are often regarded as high-stakes because without successfully completing the aforementioned assessments, a trainee is unable to move forward in his/her medical school training. Furthermore, the lack of remediation options available for trainees demonstrating deficiencies has been shown to impact raters’ willingness to provide low grades or negative evaluations even when they are warranted (Dudek, et al., 2005; Kogan, Conforti, Bernabeo, Iobst, & Holmboe, 2011). Additionally, some researchers and institutions maintain that assessments used in undergraduate medical education should serve both formative and summative purposes (e.g., Berendonk, Stalmeijer, & Schuwirth, 2013; Townsend, McLlvenny, Miller & Dunn, 2001), which is problematic when raters begin to develop personalized beliefs about how high-stakes assessments should be employed and used in order to facilitate medical trainees’ entry into practice. For example, OSCEs are commonly used to assess, at particular benchmarks, whether a medical candidate should be permitted to move forward in his/her pursuit of a medical degree. Despite this seemingly clear purpose, many raters feel that OSCEs should be used as an opportunity to provide medical trainees with feedback that would better support them in their development throughout later stages of medical training (Hill, Guinea, & McCarthy, 1994). As a rater in this context, it can be difficult to try to satisfy both the formative and summative demands of assessing medical competence without one form of assessment having the upper hand.

Within medical education it has been understood for some time that medical competence is easy to recognize but hard to measure. Medical competence is not a unidimensional construct; rather, it is a multidimensional, complex construct. At the very least, medical competence consists of two main components: clinical knowledge (i.e., the medical information) and
humanistic traits (e.g., how a person conducts him/herself and relates to patients). Currently, many of the assessments used to evaluate clinical skills are inadequate in their attempts to obtain meaningful information about medical trainees. Typically, scoring instruments are designed to assess only one component of medical competence at a time. The simplicity of some assessment instruments could be contributing to the fact that many raters struggle to translate the degree to which a performance was completed, especially in situations where a candidate carried out a clinical task in an inhumane manner. Therefore, the study conducted in this context explored the potential of an additional “attempted” checklist category to supplement the existing “done” and “not done” options in order to enhance raters’ abilities to communicate and articulate nuances and discrepancies as they make evaluative judgments. It is believed that this straightforward addition of “attempted” would afford raters the opportunity to provide judgments of both clinical competence and humanistic behaviours simultaneously.

**The Postgraduate Context**

Assessments conducted throughout one’s postgraduate medical training are valuable for not only providing feedback, but also for identifying and mentoring individual physicians in specialized areas. These assessments are not high-stakes per se, but they are critical in preparing physicians for future certification and licensure examinations, which are extremely high-stakes. Successful completion of these licensure examinations is necessary in order to enter into unsupervised clinical practice (Medical Council of Canada, 2014). Within Canada, physicians who fail their licensure examinations at the postgraduate level have invested time, money, and resources in becoming a physician, but unfortunately are unable to practice medicine independently, which means they are unable to work in a career that they have prepared for. An additional aspect that makes the postgraduate context unique is the variability associated with
various medical specialties (e.g., Surgery, Pediatrics, and Emergency Medicine). Postgraduate training is essentially about fostering one’s skills in order to become an expert in a specific area of medicine. Therefore, each postgraduate specialty requires its own set of criteria for assessing competence given that the knowledge and skills required to be successful in one specialty may not necessarily apply to another specialty or may not be equally weighted in importance.

Because of these unique challenges, assessment of medical competence within the context of postgraduate medical education continues to be a critical area of interest for researchers. Given the scope and diversity of postgraduate training, it is difficult to find an assessment that is suitable across the several different postgraduate specialties. Competency-based medical education, otherwise known as the core of postgraduate training and assessment, strives to promote fair, transparent, and defensible assessments of individuals’ medical competence (Schuwirth & van der Vleuten, 2012). As many medical schools and external organizations strive to develop assessment programs that meet the aforementioned criteria, there exists the issue of integrating quantitative assessment data with qualitative assessment data, especially in situations where the two provide divergent information. As more studies continue to highlight the shortcomings of quantitative methods for the assessment of clinical competence, there exists a greater push towards taking advantage of qualitative data such as performance reports and narrative comments. However, relying heavily on qualitative sources is not only burdensome from an analysis perspective, but harder to defend as qualitative data is implicitly viewed as being highly interpretive (Walker, 1989). Hence, the study designed for the postgraduate context aims to explore how raters’ quantitative and qualitative reports of individuals’ competence can be integrated for the purpose of obtaining a more complete impression about one’s medical competence.
Dissertation Overview

This dissertation includes a literature review (i.e., Chapter 2) that offers an overview of current research in the area and identifies some of the common challenges associated with rater-based assessments. The subsequent chapters describe empirical research from three separate studies. The first of these studies Seeing Things Differently or Seeing Different Things? Exploring Raters’ Associations of Non-Cognitive Attributes explored raters’ evaluative judgments by examining the unconscious or subconscious associations that raters make as they formulate judgments. The second study, Understanding the Shades of Grey between “Done” and “Not Done” in Objective Structured Clinical Examination (OSCE) Assessments, examined how raters judgments were affected when they were provided with an additional checklist option, removing the forced dichotomy created around competence (which is presented by the constraints of dichotomous tools). The third study, Mixed Messages or Miscommunication? Investigating the Relationship between Assessors Workplace-Based Assessment Scores and Written Comments, provides insight into the relationships that exist between raters’ quantitative and qualitative judgments.

Chapter 2 provides an overview of current research and identifies some of the common challenges associated with rater-based assessments. Rater effects are discussed to demonstrate the body of work that has previously examined the impact of raters’ behaviour in assessment situations. Then, focusing primarily on medical education literature, rater behaviour is explored according to three different perspectives (i.e., behavioural, cognitive, and sociocultural) in order to better understand why raters appear to have difficulty articulating their evaluative judgments. Together, the literature review provides a basis for investigating the questions posed in this
dissertation and offers a theoretical framework to serve as a foundation for examining the complexities associated with attempting to assess medical competence.

The first of the three studies is presented in Chapter 3. This chapter includes research exploring the insights that can be gleaned from better understanding how individual raters view relationships amongst various non-cognitive attributes. Medical admissions data obtained from a MMI used at one Canadian medical school were analyzed. This particular MMI consisted of eight stations in which candidates were rated separately by a faculty member and an upper-year medical student at each station. The Many-Facet Rasch Model (MFRM) was used to identify inconsistent raters, stations, and attributes. Hierarchical clustering with Pearson distance measures and average linkage criteria were used to examine the similarities and differences in terms of how raters viewed the various attributes measured on the MMI. This study found that having a predesigned assessment system where raters are asked to supply independent ratings for various non-cognitive competencies and attributes contributes to rater variance. It was identified that some raters were unable to discriminate between different competencies and attributes, which suggests that the way in which raters were asked to provide information hindered their ability to articulate their evaluative judgments and contributed to unexplained variance.

The research presented in Chapter 4 explores how raters use different assessment formats by investigating the impact of adding an “attempted” category in addition to traditional “done” and “not done” options. The data were obtained from an undergraduate OSCE at one Canadian medical school. This particular OSCE consisted of seven performance stations and one written station. Candidates were rated separately by a single faculty rater at each station. In order to assess the impact of the “attempted” category on psychometric aspects, the following statistical analyses (i.e., reliability measures, item-total score correlations, and ANOVAs) were completed.
Additionally, interviews occurred with select raters that were designed to assess how well raters’ felt the “attempted” category allowed them to communicate and articulate various aspects of candidates’ performances and identify situations most suited for use of the “attempted” category. This study highlighted some of the short-comings of assessment instruments and found that raters feel confined when asked to translate their judgments using existing assessment tools (e.g., checklists). Assuming that raters would make the same judgments regardless of the tool provided, suggests that the lack of options available to identify complex clinical encounters may facilitate rater variance as raters attempt to force their judgments into a category of ‘best fit’ because no category reflecting the true experience existed.

Chapter 5 addresses the alignment between various methods of assessment by examining the relationship between raters’ quantitative and qualitative judgments using data obtained from WBAs at one Canadian medical school. These WBAs were part of an established assessment program used within the institution’s Emergency Medicine postgraduate training program. Data were collected from approximately 25 raters over two consecutive years and were analyzed using logistic regression and content analysis. To further examine raters’ communication of information using both quantitative and qualitative methods, this study explored whether allowing raters to provide evaluative judgments using several different assessment instruments addresses some of the restrictions with assessment tools that raters identified in chapter 4. Even when constraints were removed and raters were provided with multiple opportunities to communicate their evaluative judgments, some raters made a conscious choice not to provide certain assessment information. Examining this phenomenon further suggests that raters lack trust in the assessment systems currently used in medical education based on the development of a hidden code, which is presently used to articulate raters “true” evaluative judgments.
Taken together, the findings from Chapters 3, 4, and 5 identify challenges with designing assessment instruments that allow raters to freely communicate their evaluative judgments. The results of this dissertation also support the idea that perhaps not all raters are equally suited for evaluating high-stakes performance assessments. Chapter 6 provides a more in-depth synthesis of the findings from the three studies and discusses some of the limitations and implications of this research. This dissertation consists of papers that have been published or are in the process of being published and, as a result, the chapters may seem awkward with separate abstracts and references for each chapter; however, each chapter represents a study that is designed to stand on its own.
CHAPTER 2: LITERATURE REVIEW AND THEORETICAL FRAMEWORK

There exists a vast amount of literature examining rater behaviour from cognitive, behavioural, and sociocultural perspectives (Gingerich, Kogan, Yeates, Govaerts, & Holmboe, 2014). Although a considerable amount of the literature specifically examining raters in the context of medical education is relatively recent, studies exploring rater behaviour in other fields (e.g., psychometrics) date back several decades. In order to better understand some of the medical education specific issues associated with raters, this chapter is going to begin with a brief summary of how rater behaviour is conceptualized. Then, using the framework outlined by Gingerich et al. (2014), recent literature investigating raters in the context of medical education will be summarized. Finally, this chapter will conclude by outlining a theoretical framework that could be used to explore why raters behave the way they do in the context of medical, high-stakes performance assessments.

Understanding the Term – “Rater Effects”

From a psychometric perspective, the idiosyncratic rating behaviour of individuals has not only been observed, but also categorized according to impact. The term “rater effects” is most commonly used to describe instances where variation within an assessment is due to some individual characteristic, perception, or interpretation on the part of a particular rater (Dobria, 2011). Myford and Wolfe (2004) classify four main categories of rater effects: leniency/severity, central tendency, restriction-of-range, and halo. The term leniency is used to describe the behaviour of individuals who consistently rate above the midpoint of a scale, while severity refers to those who rate below the midpoint of a particular scale (Myford & Wolfe, 2004). Within medical education, this phenomenon is sometimes referred to as the “hawks and doves”
syndrome (Bartman, Smee, & Roy, 2013). Hawks are individuals who provide consistently low ratings for candidates and doves are those who consistently give high ratings.

A central tendency effect occurs when a rater avoids using the outermost categories and overuses the middle categories (Linacre, 2010). Central tendency most often occurs because the rater is afraid of making a mistake. The problem that arises when raters overuse the middle categories is that every candidate rated looks average and thus, their ratings become of little value because the rater was unable to discriminate among candidates. Ultimately, when raters exhibit a central tendency effect they provide limited information about any given candidate and thereby inhibit any meaningful appraisal of a candidate’s knowledge or competence.

Range restriction, closely related to central tendency, occurs when all of the raters avoid using certain categories or have scores that are clustered together, not necessarily around the midpoint (Myford & Wolfe, 2004). One of the problems with restriction-of-range effects is that the rating scale is not fully represented, which means that the criteria against which a candidate is being compared does not align with the level of the candidate’s performance. When the assessment tools provided to raters do not match the task, raters will sometimes restrict the range to communicate and reflect the overall quality of performances they observed.

Engelhard (1994) defined the halo effect as situations in which the rater assesses a candidate holistically rather than on an item-by-item basis. Others, such as Linacre (2010), suggest that a halo effect may also occur in circumstances where an evaluation given by a previous rater influences the rating provided by the subsequent rater. Halo effects are the result of a rater who is unable to separate independent aspects of a candidate's behaviour. Other rater effects may include observational inaccuracy, contrast error, primacy error, order effects, influences of rater/applicant background characteristics, and influences of rater biases, beliefs,
attitudes, and personality characteristics, (Berendonk, Stalmeijer, & Schuwirth, 2013; Myford & Wolfe, 2004; Yeates, O'Neill, Mann, & Eva, 2012; Yeates, O'Neill, Mann, & Eva, 2013a). The main purpose of this dissertation is to better understand how and why raters exhibit these seemingly idiosyncratic behaviours when asked to provide evaluative judgments about the quality of a candidate’s performance.

When two or more raters view the same performance, they will often assign different scores based on their individual interpretations of the observed performance (e.g., Mazor, et al., 2007). Although it is believed that differences in raters’ evaluations of the same performance primarily depend upon two main factors: attending to differing aspects of performance and differing expectations regarding the standard of performance (Williams, Klamen, & McGaghie, 2003), a number of alternative explanations have been put forth to describe why raters are able to see the same performance differently. In a recent paper written by Gingerich and colleagues (2014) three different perspectives, each derived from a distinct theoretical basis, are put forth to help explain the diverse ways raters’ behaviours are viewed and understood. The following sections will describe the three rater cognition perspectives: assessors as fallible, assessors as trainable, and assessors as meaningfully idiosyncratic (Gingerich et al., 2014). Each aspect is described and augmented to include additional literature pertinent to understanding rater behaviour and cognition. It is important to note that Gingerich and colleagues (2014) use the term “assessor” as opposed to “rater” to acknowledge that raters provide assessment information that is not always categorical or numerical. For the purpose of this review, the terms assessor and rater can be viewed interchangeably to indicate an individual who is tasked with providing the evaluation of an observed performance.
Raters are Error-Prone

The first rater cognition perspective – *assessors as fallible* – uses literature from social and cognitive psychology to explain certain emotional and mental limitations that prevent raters from being able observe a performance objectively. The main thesis surrounding this perspective is that raters are not always capable of being truly accurate because they are susceptible to influences such as stereotyping or other cognitive distortions. Raters’ values, emotions, and first impressions all have the ability to manipulate the lens through which raters observe and interpret a given performance. Furthermore, the conditions (e.g., mental workload demands) surrounding many performance-based assessments can also lead to the unintentional cognitive oversights on the part of raters during evaluations of performance assessments (e.g., Tavares & Eva, 2013).

Raters’ Cherished Values

As performance assessors, raters are required to pay attention to many features, often simultaneously. Hodges (2012) suggests that performance can be viewed as a process, consisting of different features including expression, body language, and movement. Every rater will notice different aspects of a given performance, most often according to what they value, which can lead to different interpretations regarding the quality of a performance. Yeates and colleagues (2013a) found that what was important to one rater was not necessarily important to the next. In many cases, raters can articulate their cherished values and in some instances they will justify the importance of those values. Take the testimony of this one rater:

“the first thing I always pick is the interpersonal communication portion because it just happens to be 90% of what our job is. So no matter what the science or disease is – it very much comes down to how you relate to the patient” (Kogan, Conforti, Bernabeo, Iobst, & Holmboe, 2011, p. 1051).
When situations emerge that force raters to pay attention to characteristics or behaviours that are incongruent with those that raters inherently value, cognitive dissonance occurs and raters try to adjust themselves (to varying degrees) to resolve this discomfort (Festinger & Carlsmith, 1959). The problem is that raters have a difficult time separating their personal value systems from what they are asked to assess. Attributes such as communication, professionalism, and organization have become more frequently associated with medical competence, especially when it comes to obtaining licensure and certification credentials; however, the hierarchy of these competencies as well as their relation to concrete medical knowledge is often tacit. Thus, performance-based assessments are filtered through individual raters value systems in terms of what they think comprises a competent medical professional.

**Raters’ Emotional Responses**

Emotion is a term used to depict aspects such as a person’s feelings, experiences, or mental and physical states (McConnell & Eva, 2012). Emotions are connected to our thought processing; therefore, when an individual observes or interprets something that can be linked to an already formulated judgment or experience, they automatically generate an emotional response (McConnell & Eva, 2012). Emotions in the context of performance-based assessments present a serious issue as they have the potential to influence the conditions under which evaluative interpretations are made. In several instances, raters rely on their emotional responses to guide their assessment about the competence of a medical candidate. It is not uncommon to hear raters say things such as “a lot of it is just instinct” (Kogan et al., 2011, p. 1051), “I had a good feeling in what she did” Berendonk et al., 2013, p. 565), or “I’d have to get my own emotions under check…” (Kogan et al., 2012, p. 209). Gingerich, Regehr, and Eva (2011) suggest that one’s mood and emotional state can have a significant impact on raters’
interpretation of an individual during performance-based assessments. Hence, emotions play a vital role not only in raters’ judgments of medical competence, but also in the ways that raters observe performance, use assessment tools, and provide feedback (McNaughton & LeBlanc, 2012).

**Raters’ Impression Formation**

The literature surrounding impression formation has been largely divided into three areas: person models, nominal categorization, and dimensionally based categorizations (Gingerich et al., 2011). Person models focus on the emotions and values that raters use to better understand an individual of interest. Person models consider the social aspects that influence a rater’s ability to provide precise and accurate information. Nominal categorization literature examines how raters categorize information according to their previously established schemas. Nominal categorization allows raters to use short-hand methods for collecting, prioritizing, and organizing sensory information. Finally, dimensionally based categorizations involves processing information based on two independent measures, mainly cognitive attributes such as intelligence and non-cognitive attributes such as empathy or humanism. In essence, the dimensionally based categorization approach allows individuals to formulate a judgment based upon a person’s level of warmth and competence in a given area (Gingerich et al., 2011). Most recently, individuals such as Wood (2014) have been exploring the role of first impressions in raters’ ability to provide accurate interpretations during performance-based assessments. The present understanding is that raters use first impressions to help guide their evaluation of individuals (Govaerts, Van de Wiel, Schuwirth, van der Vleuten, & Muijtjens, 2013; Wood, 2014). What still remains inconclusive is the extent to which raters’ can mentally adjust and move beyond
those initial first impressions to provide dependable accounts of what occurred throughout a performance assessment.

Raters’ Mental Workload Capabilities

The cognitive demands that face a rater as they evaluate and make judgments about a given candidate can lead to variability in the rating process. The influence of cognitive aspects such as one’s cognitive capacity may serve to explain some noteworthy differences between raters (Tavares & Eva, 2013). One theory put forth to explain the cognitive limitations of raters is cognitive load theory (CLT). Originating in the 1980s, CLT suggests that individuals have limited working memory, which restricts the amount of information that can be processed. According to the principles of CLT there are three types of cognitive load (i.e., intrinsic, extraneous, and germane) that affect working memory (Sweller, Van Merrienboer & Paas, 1998). Intrinsic load refers to the difficulty or complexity of a task and germane load is essentially the coping mechanisms of working memory (e.g., the development of schemas), which handles intrinsic cognitive load. Meanwhile, extraneous load relates to external conditions imposed on an individual during a task (Van Merrienboer & Sweller, 2010). While a large body of literature exists around various aspects of CLT, Tavares and Eva (2013) attempted to re-examine the issues surrounding cognitive load using an alternative discourse that investigates the cognitive effort (i.e., the mental workload) required as raters carry out the task of assessing performance. The concept of mental workload is particularly important in medical settings where assessment often occurs in dynamic and complex environments such as outpatient clinics, hospital wards, or emergency rooms. Tavares and Eva (2013) suggest that clinical skills appraisal in this context can overload raters’ cognitive processing abilities because the cognitive demands do not always align with raters’ mental resources. One study specifically looking at the mental workload of
raters in OSCE situations suggests that one’s mental workload capacity significantly impacts some raters’ abilities to provide accurate assessment information (Byrne, Tweed, & Halligan, 2014).

**Dual Processing Memory Models**

Although not applied to rater-based assessments within the medical education literature, dual process theories of thinking have been extensively applied to rater-based judgments in other fields such as psychology. Based on the work of Epstein (1994) and others, dual process theory describes two systems in which decisions can be made. The first, sometimes referred to as System 1, is characterized by an individual’s intuitive response that is rapid and unconscious (Norman 2009; Pelaccia, Tardif, Triby, & Charlin, 2011). The other, commonly named System 2, describes more deliberate judgment that is analytic and conscious (Norman 2009; Pelaccia et al., 2011). According to Norman (2009), System 1, which relies upon an individual’s gut feeling, is not any worse in terms of being error prone than System 2, which is more deliberate and rationalized. Most of the research examining dual process memory models within medical education has examined its role in clinical reasoning; however, dual process models remain an important consideration when investigating raters’ evaluative judgments.

Taking all of this information together, it is evident that there are several probable causes of raters’ inability to provide consistent information during performance assessments. Though some of these limitations might be within raters’ control, many of these aspects that mislead raters’ are either automatic or inherent to whom they are as individuals. Perhaps a future direction could be to better understand what individual raters are subconsciously doing when evaluating and judging candidates’ medical competence by using exercises such as think alouds to make tacit assumptions and values more explicit.
Raters can be Educated

The second rater cognition perspective – *assessors as trainable* – applies a behavioural lens to justify why raters are unable to agree on various aspects of a performance. The rationale developed throughout this perspective is that raters are inappropriately interpreting and applying assessment criteria. In some cases, this may occur because raters are unfamiliar with what they are expected to evaluate. Raters’ skills and experiences are variable, which means that they will observe and regard different aspects of a performance as relevant based on their ability to synthesize and deduce information. Much of the literature around raters’ applications and interpretations of assessment criteria has focused on raters’ experiences and receptiveness to become further educated.

Differences in Raters’ Frames of Reference

Previous literature has shown that raters find it difficult to identify and articulate a standard for performance (Williams et al., 2003). In many cases, research has demonstrated that raters use their own approach to practicing medicine as a frame of reference for observing and evaluating the performance of others (Kogan, Hess, Conforti, & Holmboe, 2010). Given the personalized nature of how raters evaluate trainees during performance-based assessments, recent studies have focused on the various strategies raters use when appraising performance. Kogan and colleagues (2011) identified three common frames of reference used by raters: 1) performance based on how the individual him/herself would do it, 2) performance based on how others (typically colleagues and residents) would do it, and 3) performance based on criteria standards for patient care. Even within these commonly used frames of reference, there still exists variability because how one rater might do something could drastically differ from another
rater. To complicate this further, within medicine, although there are standards for practice, there are often multiple ways of treating a patient or achieving a desired outcome.

Other studies examining how raters develop a frame of reference suggest that raters tend to evaluate and judge performances relative to other performances they observed (Yeates, O’Neill, Mann, & Eva, 2013b). The norm-referenced, rather than criterion-referenced, approach commonly used by raters in performance appraisal is problematic because it forces raters to rely on memory, which is most likely biased and highly influenced by external factors (Yeates et al., 2013b). Yeates and colleagues (2012) found that mere exposure to the performance of a strong or weak medical candidate was enough to significantly alter the assessment of subsequent learners. Furthermore, when raters use these norm-referenced ways of making evaluative judgments no mechanisms are currently in place to determine whether raters are consistent in their development of a norm.

**Differences in Raters’ Experiences**

Since the inception of performance-based assessments, the role of raters has shifted from observer to interpreter (Regehr, MacRae, Reznick, & Szalay, 1998). As raters take on more of an interpreter role, an increase in the variability of perspectives is likely to occur because individuals tend to filter and mediate what they observe using their personal knowledge and experiences. Dreyfus and Dreyfus (1986) identified three aspects that reflect the change from novice to expert: 1) a shift from identifying abstract concepts to developing concrete examples based on lived experience, 2) viewing discrete aspects as interconnected parts of a larger whole, and 3) moving from simply observing to actively participating. Raters use different approaches in performance assessments because the processes involved in formulating evaluative judgments vary according to the individual. For instance, recent research by Govaerts, Schuwirth, van der
Vleuten, and Muijtjens (2011) showed, in the context of WBA, expert raters interpret more information and are more mindful of contextual cues than are non-experts. Other research has shown as individuals grow and gain experience, they move away from thinking and acting like a novice and become more proficient, which checklists and some rating scales are failing to detect (Hodges, Regehr, McNaughton, Tiberius, & Hanson, 1999). Without a doubt, individuals’ experience levels matter; however, it is important to keep in mind that what a novice observes and interprets can drastically differ from that of an expert.

**Differences in the Use of Scales and Instruments**

Somewhat related to the experience levels of raters is the body of literature that examines scales and instruments most commonly used in performance-based assessments. A systematic review conducted by Kogan, Holmboe, and Hauer (2009) demonstrated that the majority of tools currently used in the assessment of clinical skills among medical professionals fail to produce sufficient validity evidence to support their use. Crossley, Johnson, Booth, and Wade (2011) argue that we are unable to gather information from raters because the constructs we seek to measure are not well aligned with the scales we presently use for assessment. Their study showed that reliability increased with the development of construct-aligned assessment scales, which suggests that the inconsistencies detected psychometrically are perhaps a result of raters’ different interpretations of poorly aligned scales (Crossley et al., 2011). When assessment tools and scales are not well aligned to the construct they strive to measure, raters are forced to make interpretations in order to articulate information using these poorly aligned scales. Raters are resilient, which means that when they are provided with tools and assessment instruments, which are flawed, they adapt and often in different ways. Therefore, even in situations where raters may agree on the assessment of an individuals’ performance, it could be difficult to detect this
because the information is altered according to the different ways raters apply the assessment scales (Crossley & Jolly, 2012).

**Differences in the Effectiveness of Rater Training**

There are researchers (i.e., Green & Holmboe, 2010) who argue that the issues surrounding the evaluation of medical competence has little to do with assessment instruments and more with the variation of inadequately trained raters. Rater training is often suggested as a solution to dealing with raters' differences; however, rater training is a contentious issue within medical education because although some studies have shown satisfactory improvements with rater training (e.g., Holmboe, Hawkins, & Huot, 2004), the majority of studies suggest that rater training is ineffective (e.g., Cook, Dupras, Beckman, Thomas, & Pankratz, 2009; Williams et al., 2003). Lurie, Mooney, and Lyness (2009) suggest that rater training is ineffective because individuals have a minimal understanding of the socially constructed competencies that are claimed to be measured with performance-based assessments. However, institutions (e.g., Artemiou, Hecker, Adams, and Coe, 2015) that dedicate a substantial amount of time (i.e., 8 hours or more) to train raters on specific competencies have shown remarkable results; therefore, it is possible that rater training is seen as ineffective because raters are asked to use poorly developed tools and not given sufficient amount of time for appropriate training.

Research on the effectiveness of rater training within medical education has demonstrated that providing training does not appear to alter raters' evaluations. Hence, the behavioural perspective of raters as "trainable" is problematic because it fails to address an underlying issue: What if raters are using the scales and instruments differently as a way of communicating? Perhaps raters are trying to articulate, in different ways, similar aspects of a performance and the lack of flexibility and alignment of most scales is making it appear as though they need to be
Raters can Provide Different, but Relevant Pieces of Information

The final rater cognition perspective – *assessors as meaningfully idiosyncratic* – draws upon cognitive and socio-cultural theories to defend the inconsistencies recognized among raters. The idea behind this perspective is that the variability attributable to raters can actually be a useful source of assessment information. Raters, through their personalized idiosyncrasies, have unique and distinct interpretations of a performance and it is through the process of embracing those different interpretations that we can achieve a more complete (and potentially accurate) representation of a given performance. Nevertheless, the viewpoint that is currently being argued asserts that to view raters as a main source of error is detrimental to capturing the valuable information that raters are capable of producing (Govaerts & van der Vleuten, 2013). Thus, focusing on rater errors makes it difficult to ascertain what raters are able to do well.

Different Ways of Understanding Rater Idiosyncrasy

For several decades, researchers believed that inquiry into the cognitive aspects of raters was the best way to understand and explain how individuals engage with the rating process (Landy & Farr, 1980). However, more recently researchers have come to acknowledge that other non-cognitive features also play a role. Yeates et al. (2013a) identified three positions that can be used to describe how variability originates among raters: 1) raters value different things and what they pay attention to during a performance often reflects those values, 2) in line with Kogan et al.’s (2011) findings, raters use different reference criteria, which generates variability in assessments, and 3) raters attitudes and emotions, even if they are positive or negative, are used to develop of a holistic impression regarding one’s competence. These findings suggest raters rely on different, individualized combinations of information when interpreting a performance.
Furthermore, current psychometric approaches are failing to detect these individualized perspectives, and thus, these meaningful differences among raters are unequivocally labelled as rater effects or error.

Most recently, Berendonk and colleagues (2013) decided to investigate variation from the perspective of raters and found three main categories for how they perceive themselves making decisions. Firstly, they identified that all raters had insecurities about assessment (Berendonk et al., 2013). Many raters articulated that they did not feel they had sufficient content knowledge or experience with trainees at the level in which they were assessing. Secondly, raters’ perceptions and beliefs about the nature, purpose, and value of assessments seemed to vary (Berendonk et al., 2013). Presently, many assessments are being used for both formative and summative purposes (Berendonk et al., 2013; Townsend et al., 2001), which likely creates misconceptions about raters’ roles as performance assessors. Finally, the assessment of trainees occurs within a particular context. Some raters lack confidence in their institution’s procedures and regulations for assessment and remediation (see Dudek, Marks, & Regehr, 2005), while others simply fail to buy-in to certain forms of assessment (e.g., OSCEs), both of which may be responsible for raters’ idiosyncratic behaviour.

**Variation due to Different Perceptions of the Same Performance**

Given the individuality associated with raters, it is important to consider how they synthesize information to make evaluative judgments. According to Williams and colleagues (2003), raters’ capacities to formulate an impression about one’s clinical performance are limited to one or two dimensions, most likely knowledge and soft skills (e.g., professionalism). However, other research in this area demonstrates that based on the individual being evaluated, raters can idiosyncratically apply varying amounts of importance across eight different
dimensions that include knowledge, professionalism, patient interactions, team interactions, systems, disposition, trust, and impact on staff (Ginsburg, McIlroy, Oulanova, Eva, & Regehr, 2010). Further, studies in this area have recently argued that observed halo effects are in fact due to raters evaluating performance using several different dimensions (Govaerts et al., 2013; cf. Murphy & Reynolds, 1988). The identified variability in assessments stemming from raters’ perceived importance of various dimensions introduces a new area of research focused on how to capture, with a certain degree of precision, crucial evaluative information during performance-based assessments.

The “assessors as meaningfully idiosyncratic” perspective has opened up the door for individuals to view these different interpretations as “equally valid” (e.g. Landy & Farr, 1980; Yeates et al., 2013a). Although different interpretations can provide a more holistic picture of a given performance if raters are using different dimensions, and are variable in the importance of those dimensions, then those interpretations cannot be equal because not all dimensions hold the same value in a given context. Therefore, the future of performance-based assessments relies on knowing when raters are privileging certain dimensions so that judgments about the quality of these different interpretations can be ascertained.

**Conclusion**

Assessment is a human process, conducted by and with human beings, and subject inevitably to the frailties of human judgment. However, crisp and objective we try to make it, and however neatly quantifiable may be our ‘results’, assessment is closer to an art than a science. It is after all, an exercise in human communication (Sutton, 1992, p. 2)

The recent desire to deliver competency-based medical education has not only emphasized the importance of clearly defining competence, but has also generated debates as to the best practices for teaching and measuring medical competence. Currently, the assessment of clinical competence occurs through the use of performance-based assessments such as OSCEs
and WBAs. However, these assessments can be problematic as they rely on the judgments of qualified medical professionals as designated assessors of clinical competence. According to existing literature, raters employed in performance-based assessments are idiosyncratic and have been shown to provide inconsistent, and sometimes irrelevant, information in their evaluative judgments (see Gingerich et al., 2011).

Although multiple perspectives encompassing behavioural, social, and sociocultural theories, have been put forth to explain the idiosyncratic behaviours associated with raters, there exists an underlying problem of communication and articulation in all of those perspectives. Communication is a crucial aspect of assessment because regardless of the accuracy of raters’ judgments if raters are unable to articulate those judgments, then the information obtained from raters is worthless. Therefore, the best approach for eventually being able to accurately assess medical competence, through the use of raters in performance assessments, is to gather insight into how raters communicate and articulate their evaluative judgments, especially when they are provided with flawed assessment tools. The fact of the matter is that although we have ideas about what raters are doing, we still do not know when and why raters are exhibiting such idiosyncratic behaviours. The future of assessing competence within medical education depends upon knowing what raters are doing, and when they do it, as well as why raters are behaving the way they do. In summary, this research program aims to better understand how raters communicate and articulate what they see, interpret, and evaluate during the assessment process.
Theoretical Framework

*Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality.*

- Thorndike, 1918, p. 16

To better understand how raters articulate their judgments about a particular performance, this dissertation explored the process of impression formation and interpretative judgments using the dimensionally based categorization perspective (Gingerich et al., 2011). This particular model was selected to serve as the theoretical framework because it considers judgments based on dimensional scales (Gingerich et al., 2011). Literature focusing on this form of categorization has identified that judgments are largely based along two independent measures that are orthogonally oriented in space (Smith & Medin, 1981). The first measure (dimension) represents the more traditional view of competence, which focuses on an individual’s knowledge and abilities in a particular area such as history taking or physical examinations in medical situations (Williams et al., 2003). Meanwhile, the second measure usually relates to more contemporary aspects of competence, which encompass “soft skills” and other non-cognitive attributes (e.g., communication, humanism, or professionalism), which have more recently become greater emphasized and valued in the assessment of medical competence (Schuwirth & van der Vleuten, 2012). The intersection of these two dimensions results in four possible quadrants. Literature has shown that raters use the two different dimensions to categorize an individual’s performance (Gingerich et al., 2011). Raters make evaluative judgments about an individual’s overall competence by placing them somewhere within one of the four quadrants (see Figure 2.1). From a dimensionally based categorization perspective, problems arise when an individual is placed near the cusp of a dimension or in a position that is equally spaced between the two dimensions.
where one measure is high and the other is low as these instances both represent unclear cases (Smith & Medin, 1981).

![Figure 2.1. The Four Quadrants Derived from the Dimensionally Based Categorization Perspective](image)

The idea that individuals perceive things and formulate judgments according to measures along two separate dimensions suggests that any assessments involving raters need to consider both dimensional aspects. In assessment situations where the two aspects are not considered, and raters are asked to provide a holistic judgment (e.g., competent or not), cognitive dissonance may occur. Cognitive dissonance refers to personal discomfort arising from inconsistencies in one’s opinions or attitudes, for instance, simultaneously holding two or more contradictory beliefs, ideas, or impressions (Festinger, 1957). According to Festinger (1957), when an individual experiences cognitive dissonance, the discomfort will motivate the individual to actively reduce
the inconsistency. For example, when a medical trainee asks all the right questions while taking a history, but does so in a way that feels like an interrogation, it can be difficult to conceive that someone factually competent can lack certain humanistic qualities. A rater assessing this situation might overlook the humanistic aspect and award the trainee a satisfactory assessment. Cognitive dissonance has been shown to result in individuals’ defaulting to judgments based on how much they liked the individual being assessed rather than his/her ability to perform a particular task. In other instances, cognitive dissonance has caused individuals make an evaluative judgment based on some idiosyncratic scale of reference (Upshaw, 1968). When trainees are high or low on both dimensions used for the basis of categorization, they are easier for raters to identify and evaluate because of the consistency across the two measures. An interesting challenge for raters occurs in those difficult, unclear cases where individuals are high on one dimension and low on the other because the process of resolving such inconsistencies in performance-based assessments are rarely explicitly described.

Designing assessments in which the construct aligns with the mental processes of raters is confounded by raters’ cognitive dissonance in situations where trainees’ observed behaviours are inconsistent. These specific instances may be contributing to the CIV observed in rater-based assessments (Dudek et al., 2005; Fazio, Papp, Torre, & Defer, 2013; Kogan et al., 2011). It appears that raters, when asked to make evaluative judgments that are challenging or ambiguous, consistently overestimate an individual’s level of performance because they give the individual being assessed the benefit of the doubt (Dhar & Simonson, 2003). Evidence of this impact was observed in a recent study in which the findings suggested that almost 40% of medical trainees passed when they should have failed based on the judgment of a trustworthy assessor (Fazio et al., 2013). Raters’ abilities to resolve cognitive dissonance is especially problematic in
situations where raters are being asked to provide judgments based on discrete categories because these measures have the potential to limit how well raters are able communicate important assessment information (e.g., precision in the quality of abilities and attributes assessed). An inability to adequately obtain rigorous evaluations, especially in high-stakes assessment situations, could be responsible for a loss of crucial information that never actually gets communicated. Therefore, there exists a need to investigate how assessments could be used to identify and articulate meaningfully relevant information from raters in a way that encourages their subjective judgments and acknowledges any inconsistencies without compromising the accuracy of interpretations that are generated based upon trainees’ achievement.

In the context of performance-based medical assessments, achievement information obtained from raters tends to inform a holistic judgment about a medical trainee’s overall competence, and is frequently used to make high-stakes decisions (e.g., whether the candidate should pass or fail an examination or move on to the next level of medical training). Raters’ personal judgments can be problematic because each rater provides a judgment based on what he/she observed and interpreted in the context of a specific clinical encounter and “the whole is not the same as the sum of its parts” (Barnes, 1985, p. 252, as cited in Vega, Hernandez, & Rivaud, 2002, p. 83). Therefore, collecting meaningfully relevant assessment information from raters helps ensure accuracy and precision not only in raters’ evaluative judgments, but also in final decision-making processes.

Given the widespread use of rater-based assessments in medical education, and the many challenges associated with rater-mediated assessments, further research is required to understand whether current assessment practices are yielding accurate information about an individual’s level of competence. In order to examine this question in greater depth, this dissertation explores
what information is acquired from raters by investigating how raters communicate information and examining their behaviours in different performance-based assessment contexts.
CHAPTER 3: SEEING THINGS DIFFERENTLY OR SEEING DIFFERENT THINGS? EXPLORING RATERS’ ASSOCIATIONS OF NON-COGNITIVE ATTRIBUTES

Abstract

Purpose: Raters represent a significant source of unexplained, and often undesired, variance in performance-based assessments. To better understand rater variance, this study investigated how various raters, observing the same performance, perceived relationships amongst different non-cognitive attributes measured in performance assessments.

Methods: Medical admissions data from a Multiple Mini-Interview (MMI) used at one Canadian medical school were collected and subsequently analyzed using the Many Facet Rasch Model (MFRM) and hierarchical clustering. This particular MMI consisted of eight stations. At each station a faculty member and an upper year medical student rated applicants on various non-cognitive attributes including communication, critical thinking, effectiveness, empathy, integrity, maturity, professionalism, and resolution.

Results: The Rasch analyses revealed differences between faculty and student raters across the eight different MMI stations. These analyses also identified that, at times, raters were unable to distinguish between the various non-cognitive attributes. Hierarchical clustering highlighted differences in how faculty and student raters observed the various non-cognitive attributes. Additionally, differences in how individual raters associated the various attributes within a station were also observed.

Conclusions: The MFRM and hierarchical clustering helped to explain some of the variability associated with raters in a way that other measurement models are unable to capture. Our findings highlight that differences in ratings may result from raters forming different interpretations of an observed performance. This study has implications for developing more purposeful rater selection and rater profiling in performance-based assessments.
Seeing Things Differently or Seeing Different Things? Exploring Raters’ Associations of Non-Cognitive Attributes

It is well established within medical education that raters represent a significant, and often undesired, source of variance.\(^1\) Previous research has shown that raters’ scores and evaluations during performance assessments can vary drastically.\(^2\)-\(^4\) For example, Holmboe et al.\(^5\) showed that multiple raters viewing the same performance produced scores that varied by as much as 6 points on a 9-point scale. Other studies have not only observed variance in raters’ assessment of specific competencies, but also in their global assessment of individuals’ performance.\(^6\) These findings of rater inconsistency undermine confidence in performance assessments.

Variability among raters is often identified when inter-rater reliability statistics are low. There exist two types of variance: 1) systematic variance, which is the variance that represents meaningful differences due to some form of manipulation or treatment condition, and 2) error variance, which is variance often attributed to random or extraneous variables. Error variance caused by raters’ personal characteristics or differences in raters’ interpretations of how to assess competence is problematic because it can impact the validity of performance-based assessments. While case specificity also impacts the reliability of assessment scores,\(^7\) recent studies focusing on the assessment of specific competencies (rather than tasks) suggests that rater variance has a significant influence on the precision of assessment scores.\(^8\) In an attempt to address the problem of this undesired rater variance, several studies have explored rater training as an approach to increase consistency among raters.\(^9\) While some studies have shown satisfactory improvements with rater training,\(^10\) the vast majority of studies in medical education suggest that rater training is ineffective.\(^11,12\) The inability of rater training to adequately address the issue of low inter-rater
reliability has prompted researchers to explore the problem of unexplained rater variance from various practical and theoretical perspectives.

Rater variability has many potential sources. For example, Yeates and colleagues\textsuperscript{13} suggest rater differences occur because raters value different aspects of a given performance (often to varying degrees), raters are eclectic in their use of assessment criteria, and raters tend to formulate global impressions based on their own personal categorizations of events. Rater characteristics, perceptions of the assessment task, and differences in the assessment context have also been shown to impact raters’ abilities to produce consistent scores.\textsuperscript{14}

Recently, Gingerich et al.\textsuperscript{15} explored the fundamental notion of rater variability using behavioural, social, cognitive, and cultural lenses. Their theoretical approach to understanding rater variance offered three unique perspectives (i.e., assessors as trainable, assessors as fallible, and assessors as meaningfully idiosyncratic) to explain why raters vary in their assessments of performance. One perspective in particular – \textit{assessors as meaningfully idiosyncratic} – suggests that rater variance may be a result of raters formulating different, but accurate interpretations in their assessments of a particular performance. This perspective is further supported by studies showing how social and cognitive factors interact in various ways to allow for multiple interpretations of the same performance.\textsuperscript{16}

Collectively, these studies highlight several reasons why raters differ in their assessment of the same performance. Further, they also raise the possibility that unexplained variance, commonly named rater error or bias, may actually represent meaningful variability in performance, which is differentially observed by the raters. Nevertheless, if raters provide different interpretations of the same performance, then it becomes important to determine if those differences represent meaningful information as oppose to errors in judgment. As the field
of medical education continues to move towards a competency-based framework for assessing performance, understanding how raters associate various non-cognitive attributes is imperative to determining how they arrive at an interpretation of one’s overall competence. Furthermore, quantifiable forms of measurement, that can be used to obtain statistical information about raters, need to be evaluated to determine the extent to which they offer meaningful information about raters’ interpretations of non-cognitive attributes and general competence. Our research investigated how raters distinguish between different non-cognitive attributes in the context of performance-based assessments to better understand unexplained rater variance. The following research questions were addressed: 1) How can the Many Facet Rasch Model (MFRM) and hierarchical clustering be used to understand some of the unexplained variance associated with raters? 2) What insights can be gained from better understanding how individual raters view relationships amongst various non-cognitive attributes?

**Methods**

**Sample**

This research study used Multiple Mini-Interview (MMI) admissions data from 455 medical school applicants who were shortlisted for admission at one Canadian medical school. The specific focus of the individual MMI cases and the specific attributes measured at each of the MMI stations can be found in Table 3.1. This MMI consisted of eight stations, each 10 minutes in length (i.e., 8 minutes for the station and 2 minutes to complete the score sheet), where both a faculty member and an upper year medical student rated applicants separately on various combinations of non-cognitive attributes. There were three separate MMI tracks operating concurrently, which correspond to the six raters (i.e., three faculty and three students) observing the same MMI case scenario, albeit for different sets of applicants.
Rater training. All of the raters were given technical training and instructions by the director of the admissions committee. In some cases, where the student rater was also functioning as a standardized patient, additional training was provided to ensure that all the student raters were portraying the same role. All of the raters participated in rater training that covered: 1) the general purpose of the MMI, 2) the content applicants are assessed for, and 3) instructions for how to use the rating form, which included explanations for how the various attributes were measured against the scale. Finally, all six of the faculty and student raters assigned to the same station were grouped together to further discuss the information as it related to the specifics of their station.

Instruments and Procedure

All medical school applicants were assessed on a different combination of non-cognitive attributes (i.e., communication, critical thinking, effectiveness, empathy, integrity, maturity, professionalism, and resolution) within a given station. Applicants were rated using a 9-point scale for each attribute of interest. The 9-point scale was anchored at three points (i.e., 1 = ‘Poor’, 5 = ‘Satisfactory’, 9 = ‘Outstanding’). Two raters observed and scored each applicant. Thus, each prospective medical student received two sets of scores: one from a faculty member (with an existing appointment) and another from an upper-year medical student (in good standing) for each of the attributes measured within the station.

Rasch analyses. Rasch analyses were conducted on the complete set of 455 applicants using FACETS software, version 3.68.1. At each station, applicants were scored on one to four attributes. The MFRM is an extension of the original Rasch measurement model that goes beyond person ability and item difficulty to measure other factors (e.g., rater severity) which can interact with an assessment situation. The particular Rasch model used in this study was a three-
facet rating scale model, which examined applicants’ ability, raters’ severity, and attributes’ level of difficulty. Ratings obtained from the faculty and upper-year medical students were analyzed using the MFRM to obtain reliability coefficients for applicants, raters, and attributes.

Reliability coefficients were calculated separately for the three facets (i.e., applicants, raters, and attributes) within each of the stations. These reliability values represent the Rasch reliability of separation. The Rasch reliability of separation ranges from 0 to 1.0 where 0 indicates no significant differences between raters. Thus, the closer the Rasch reliability coefficients are to zero, the more confidence we have in the results.

**Hierarchical Clustering.** To identify raters’ unique interpretations, we used hierarchical clustering to cluster the non-cognitive attributes measured within each station. This procedure was carried out using R, version 3.1. The clustering phase of this study produced visual representations that model similarities between the various attributes. Each cluster represents a subgroup of the attributes that rater’s observed to be similar within their station. To calculate the dissimilarity between attributes, we used the Pearson distance calculation, which is a transformation of the Pearson correlation. Pearson distance measures are well-established in generating clusters that represent high levels of accuracy (see Fulekar19 for the calculations involved in the generation of the Pearson distance matrix). We then used hierarchical clustering to cluster individual station attributes into groups of attributes that raters interpreted to be similar. Hierarchical clustering is an iterative process, which means that each attribute starts in its own cluster and the hierarchical clustering process iteratively identifies and merges clusters that contain similar attributes. Clusters are said to contain similar attributes if the distance (as calculated by the Pearson distance) between the clusters is small. Given the relationship between the Pearson distance and Pearson correlations, Pearson distance measures of 0.20 or smaller
would indicate high degree of similarity among attributes. Unlike several other clustering approaches, hierarchical clustering does not require the number of clusters to be specified prior to performing clustering, which makes hierarchical clustering well suited as an exploratory analysis technique. Finally, dendrograms were produced to visually display relationships among attributes as interpreted by the individual raters. Dendrograms for two stations (Stations 4 and 6) were produced to provide specific examples of how information obtained from the MFRM can be used to support hierarchical clustering procedures.

Results

The information presented in Table 3.2 summarizes the Rasch reliability coefficients for each of the eight stations. Applicant reliability values ranged from 0.84 to 0.92, which suggests heterogeneity within the sample. These higher reliability values indicate larger variation (separation) in the ability levels of the medical school applicants, which is desired because it helps to increase assessment accuracy when using an item response theory model of measurement. Similarly, attribute reliabilities should also be high because this indicates the different non-cognitive attributes were being distinctly measured within a given station.

In contrast to applicant and attribute reliabilities, rater reliability coefficients should be low within the stations because this would suggest little variance among the raters. Upon investigation of individual stations, we found that Station 4 had a Rasch reliability of separation value of 0.94, which denotes substantial differences between the faculty and student raters at Station 4. Station 4 assessed effectiveness, empathy, professionalism, and resolution. Conversely, Station 6 produced a reliability coefficient of 0, suggesting there were no differences between raters at this station. Thus, Station 6 represents optimal conditions for reliability: high for applicants, high for items, and low for raters. Stations 6 assessed critical thinking,
communication, and empathy.

The hierarchical clustering yielded dendrograms, visual representations that illustrate raters’ interpretations of the relationship between various attributes (see Figures 3.1-3.4). Using two different examples, we highlight how raters (both as groups and individually) interpreted relationships between the non-cognitive attributes at their station. Figure 3.1 shows a dendrogram comparing the faculty and student raters that assessed applicants at Station 4. In total there were three faculty and three student raters at this station because of the three different tracks. The case scenario at Station 4 measured the following attributes: effectiveness, empathy, professionalism, and resolution. As noted above, Station 4 was problematic as it had a high reliability of separation value for raters (0.94). Clustering the raters at Station 4 (see Figure 3.1) revealed that faculty seemed to view professionalism as the most distinct attribute, while the upper-year medical students saw empathy as the most distinct. Additionally, the placement of attribute clusters against the centre scale (which in this instance is approximately 0.52) indicates an overall greater level of dissimilarity and higher variability at this station.

The case scenario at Station 6 measured the following three attributes: communication, critical thinking, and empathy. Station 6 represented ideal psychometric values for this type of measurement situation due to the high reliability values for applicants (0.92) and attributes (0.91), but low values for raters (0.00). Clustering the raters at Station 6 (see Figure 3.2) cross-validated the MFRM findings as we found no differences between the raters. The dendrograms for both faculty and student raters show the non-cognitive attribute of critical thinking as more distinct from communication and empathy. Furthermore, the placement of the attribute clusters against the center scale, which in this case is approximately 0.30, indicates overall lower levels of dissimilarity and less variability within this station compared to Station 4 (see Figure 3.1).
Dendrograms were then produced for individual raters. Figure 3.3 highlights the dendrograms produced for one set of faculty and student raters who shared an examination room and observed all of the same performances at Station 4 simultaneously. The representations depicted in Figure 3.3 confirm that although two raters viewed the same performance, their interpretations of the attributes captured in that performance varied. The faculty member on the first track at Station 4 viewed effectiveness and empathy as somewhat similar and professionalism and resolution as similar, which indicates that the two clusters of attributes were seen as quite distinct. Meanwhile, the student rater on the first track at Station 4 viewed the non-cognitive attributes more distinctly from each other, with empathy being the most distinct. Further, the interpretation of non-cognitive attributes by the student rater on Track 1 and Station 4 was consistent with the other student raters assigned to Station 4 (see Figure 3.1).

Finally, the dendrograms featured in Figure 3.4 represent the clustering of attributes as perceived by the faculty and student raters who viewed the same performances on Track 2 at Station 6. As discussed earlier, there were no significant differences between the faculty and student raters at Station 6 in terms of variability and how they viewed relationships among the non-cognitive attributes. The dendrograms indicate that communication and empathy were most similar to each other and critical thinking was the most distinct non-cognitive attribute measured at Station 6. Thus the faculty and student raters on Track 2 at Station 6 were consistent in their interpretations of the relationships among various attributes (i.e., communication, critical thinking, and empathy) used to evaluate medical school applicants. Although figures for Track 1 and Track 3 were not included for Station 6, the variability among the non-cognitive attributes was similarly observed for both faculty and student raters.
Discussion

The results from the MFRM and hierarchical clustering present a foundational mechanism that helps explain some of the variability associated with raters in ways other measurement models are currently unable to capture. Through application of the methods used in this study, we demonstrated that variance in ratings could be due to raters formulating different interpretations of non-cognitive attributes or using alternative ways of communicating information about the attributes we intend to measure in performance assessments. This combination of the MFRM and hierarchical clustering offers an initial opportunity to better understand the underlying structure of individuals’ rating behaviours, which could be used to further examine idiosyncratic differences in raters.

This study builds upon previous literature that seeks to advance our understandings about the nature of variance in rater-mediated assessments. Results from this study suggest that variance is not necessarily synonymous with error or bias, but rather differences. The hierarchical clustering illustrated differences in how raters viewed and interpreted relationships between various non-cognitive attributes. Although, conceptually, the notion of variance can be interpreted as positive or negative, we argue that variance is only problematic in situations where its source is unknown. Studies such as this one that focus on trying to better understand variance rather than determining its worth provide us with an opportunity to gain insights into raters’ interpretations. Raters see different things when viewing performances and idiosyncratic judgments could be their personalized way of expressing different aspects of the same performance. However, it is essential that such differences are explained and accounted for to uphold the validity of raters’ interpretations and judgments in performance assessments.

While we have focused on variability in faculty and student groups as well as individual
raters’ associations of certain non-cognitive attributes, it is also important to consider the role of variance in the broader context. For example, further consideration regarding the purpose of employing multiple raters in performance assessments is crucial. Organizations and institutions need to evaluate whether multiple raters are merely a way of justifying decisions on the basis of consensus or whether we truly believe that multiple raters offer different interpretations and perspectives that could facilitate accuracy in decision-making. If the latter is desired, then traditional measurement models based on averaging raters’ judgments may prove to be insufficient, especially in cases where there are high levels of disagreement. Therefore, adopting approaches that use multiple methods (e.g., MRFM and hierarchical clustering) to examine complex constructs should help to better identify meaningful differences in raters’ interpretations.

In the context of medical admissions, the MMI has become an increasingly popular way of screening and selecting prospective medical students; however, the MMI is not exempt from the challenges associated with rater-mediated assessments. Studies such as this one can improve the assessment process by continually monitoring individual raters as a unique source of variability in performance assessments. Being able to identify rater differences is most profitable in being able to support them at the institutional level. For instance, results from this study were used to inform future assessment practices and enhance rater training by providing raters with descriptors for each rating scale point in addition to holistic examples to look for when assessing the different attributes of medical applicants. Understanding strengths and limitations of raters could provide opportunities to better support raters when they make judgments about individuals’ competence.
Conclusions

Our research explored how two existing measurement models (i.e., the MFRM and hierarchical clustering) could be used to identify patterns in the variability that exists among raters. If these patterns represent systematic variability, then it could account for some of the unexplained variance associated with raters by identifying when different interpretations are a result of raters seeing different aspects of performance. We found that variability in raters’ assessments can result from differences in how raters associate, interpret, and evaluate various non-cognitive attributes. These findings highlight perceptual differences found in individual raters that could be used for more purposeful rater selection and rater profiling. Understanding how individual raters view relationships between various non-cognitive attributes can inform our ability to determine whether rater variance signifies differences that represent valid interpretations or simply errors in judgment.


Acknowledgments:
Peter Hugh MacDonald, MD and Rebecca Peterson-Josza, for sharing their knowledge and experience with the medical admissions process.

Funding/Support:
This study was funded by the Social Sciences and Humanities Research Council (SSHRC).

Other disclosures:
None.

Ethical approval:
This project was exempt from research ethics board approval because the study relied exclusively on secondary use of anonymous information and dissemination of the results does not generate any identifiable information.

Disclaimers:
None.

Previous presentations:
Findings from this study were presented at 2014 AAMC Medical Education Meeting, November 6, 2014, Chicago, Illinois, United States of America.
Table 3.1

**Content and Attributes for MMI Stations**

<table>
<thead>
<tr>
<th>Station #</th>
<th>Station Topic</th>
<th>Attributes Measured at the Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conversation with High School Guidance Counsellor</td>
<td>Communication, Critical Thinking, Maturity</td>
</tr>
<tr>
<td>2</td>
<td>“Commitment to Service” Presentation</td>
<td>Communication, Critical Thinking</td>
</tr>
<tr>
<td>3</td>
<td>International Volunteering Challenges</td>
<td>Communication, Critical Thinking, Maturity</td>
</tr>
<tr>
<td>4</td>
<td>Coworker’s Abusive Relationship Disclosure</td>
<td>Effectiveness, Empathy, Professionalism, Resolution</td>
</tr>
<tr>
<td>5</td>
<td>Group Assignment Involving Plagiarism</td>
<td>Integrity, Maturity</td>
</tr>
<tr>
<td>6</td>
<td>Discussion of Overseas Healthcare Treatment</td>
<td>Communication, Empathy, Critical Thinking</td>
</tr>
<tr>
<td>7</td>
<td>Rate Your Doctor’s Performance</td>
<td>Critical Thinking, Professionalism, Maturity</td>
</tr>
<tr>
<td>8</td>
<td>Panel Interview</td>
<td>Communication</td>
</tr>
</tbody>
</table>

Table 3.2

**Rasch Reliability Coefficients for Applicants, Raters, and Attributes according to Station**

<table>
<thead>
<tr>
<th>Station</th>
<th>Applicant Reliability Values</th>
<th>Attribute Reliability Values</th>
<th>Rater Reliability Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>s₁</td>
<td>0.90</td>
<td>0.76</td>
<td>0.80</td>
</tr>
<tr>
<td>s₂</td>
<td>0.92</td>
<td>0.97</td>
<td>0.60</td>
</tr>
<tr>
<td>s₃</td>
<td>0.91</td>
<td>0.07</td>
<td>0.64</td>
</tr>
<tr>
<td>s₄</td>
<td>0.89</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>s₅</td>
<td>0.89</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>s₆</td>
<td>0.92</td>
<td>0.91</td>
<td>0.00</td>
</tr>
<tr>
<td>s₇</td>
<td>0.90</td>
<td>0.93</td>
<td>0.31</td>
</tr>
<tr>
<td>s₈</td>
<td>0.84</td>
<td>N/A</td>
<td>0.61</td>
</tr>
</tbody>
</table>

*Note. Table adapted from: Psychometric properties of the multiple mini-interview used for medical admissions: Findings from generalizability and rasch analyses³*
Figure 3.1. Dendrogram for the comparison of faculty and student raters and Station 4. The height at which two attributes are joined represents how dissimilar they are. For example, from the dendrogram on the left (Station 4 - Faculty), we find that Effectiveness and Resolution are the most similar (they are "joined" at the lowest height) while Professionalism is the most distinct attribute (it is "joined" at the greatest height).

Figure 3.2. Dendrogram for the comparison of faculty and student raters at Station 6
Figure 3.3. Dendrogram comparing the faculty and student rater that simultaneously observed and rated applicants at Station 4.

Figure 3.4. Dendrogram comparing the faculty and student rater that simultaneously observed and rated applicants at station 6.
CHAPTER 4: UNDERSTANDING THE SHADES OF GREY BETWEEN “DONE” AND “NOT DONE” IN OBJECTIVE STRUCTURED CLINICAL EXAMINATION (OSCE) ASSESSMENTS

Abstract

Purpose: Raters often struggle with traditional binary checklists partially because performance tasks are falsely portrayed as dichotomous. This study removed the artificial dichotomy created by binary checklists and presented raters with an additional checklist category, which was labelled “attempted,” and assessed the impact on raters’ evaluations.

Methods: Objective structured clinical examination (OSCE) data from one undergraduate medical school were collected and analyzed to examine the psychometric impact of an “attempted” category. Subsequently, OSCE raters provided qualitative feedback and were interviewed regarding their interpretation and use of the “attempted” category.

Results: The quantitative analyses revealed differences between raters who used the traditional binary checklist and raters who used the three-point checklist. Further analyses also illustrated that the weighting of the “attempted” category can significantly impact the scores and pass rates of medical candidates. Qualitative analyses identified that raters tended to use the “attempted” category when something unusual happened, when candidates verbalized a task, but did not actually perform it, or when a candidate attempted to complete a task, but it was not done completely or correctly.

Conclusions: Using a three-point checklist, which included an “attempted” category, allowed raters an opportunity to identify situations when it was difficult to make a definitive evaluation of a particular task. Findings illustrate that the use of an additional category can impact psychometric factors such as pass rates. Overall, raters felt that the “attempted” option supported a more accurate and fair assessment of medical candidates.
Understanding the Shades of Grey between “Done” and “Not Done” in Objective Structured Clinical Examination (OSCE) Assessments

The ability to effectively articulate what raters observe and communicate is vital to the assessment of medical competence. The goal of the objective structured clinical examination (OSCE), as well as many other performance-based assessments, is to provide individuals with an opportunity to demonstrate their skills in a close-to-realistic environment (Hodges, 2006). Since the inception of OSCEs, researchers have noted that the role of raters has shifted from observer to interpreter (Regehr, MacRae, Reznick, & Szalay, 1998), yet few changes in the development of assessment tools have supported this transition. According to Crossley, Johnson, Booth and Wade (2011), attempts by researchers to modify and enhance the rating scales and tools used in performance assessments have been largely unsuccessful in eliminating variability caused by raters. As raters take on more of an interpreter role, assessment scales need to be designed in ways that align with not only measurement constructs of interest, but also the mental language raters’ use while making evaluative judgments (Crossley et al., 2011).

Raters are not always confident in their assessment of abilities (Dudek, Marks, & Regehr, 2005; Fazio, Papp, Torre, & DeFer, 2013; Kogan, Conforti, Bernabeo, Iobst, & Holmboe, 2011), resulting in a greater likelihood of error judgments on the part of raters. These incorrect judgments are typically referred to as Type I (false positive) and Type II (false negative) errors. Most often, when raters are asked to make difficult evaluative judgments, they consistently overestimate an examinee’s level of performance because they want to give the individual being assessed the benefit of the doubt (Dhar & Simonson, 2003). High-stakes assessments such as OSCEs exacerbate this problem as an incorrect decision will have important consequences for the examinee. A recent study in the United States suggests that almost 40% of individuals passed their clerkship when they should have failed (Fazio et al., 2013), which suggests that the
frequency of incorrect decisions may be occurring more often than expected. Furthermore, raters play a vital role in high-stakes assessments given that their individual judgments are often used as a basis for making pass/fail decisions.

In an attempt to reduce the errors that raters make when estimating an individual’s performance, many rating systems make use of dichotomous checklists that break down a performance task into a series of smaller tasks that can be evaluated and classified as “done” or “not done.” While identifying specific aspects of practice that are important to observe, checklists also present tasks as false dichotomies. According to the testimony of one rater, “Even just checking ‘done’ or ‘not done’, it is difficult… you also have to check sometimes on what level it has been done: good, bad, or intermediate” (Interviewee 1, as cited in Berendonk, Stalmeijer, & Schuwirth, 2013, p. 564). Not allowing raters to assess the extent to which a task was fully completed can affect the frequency of Type I and Type II errors and further exemplify the high-stakes nature of an assessment.

The literature surrounding the use of checklists for assessment purposes is wide-spread. From a psychometric perspective, research has illustrated global rating scales anchored by checklists produce more reliable measures than the use of checklists alone in the evaluation of performance in OSCE assessments (Regehr et al., 1998). Literature has also demonstrated that traditional binary OSCE checklists are less accurate among individuals that possess higher levels of competence (Hodges, Regehr, McNaughton, Tiberius, & Hanson, 1999), not to mention they also fail to capture complex aspects of performance (McKinley et al., 2008). Additionally, a systematic review examining the use of checklists for assessing clinical skills revealed that the overall quality of checklists in medical education is generally not high (McKinley et al., 2008). However, most recent research surrounding the use of checklists highlights the benefits for
novice and inexperienced raters (Sibbald, De Bruin, & van Merrienboer, 2014), examines differences between clinical discriminating and thoroughness checklists, and advocates for a more integrative approach in which raters are part of checklist development (Schumtz, Eppich, Hoffmann, Heimberg, & Manser, 2014). These ongoing findings suggest that although medical education has struggled with raters’ use of checklists to evaluate clinical competence, there still exists a desire to continue to use checklists within certain situations as they provide raters with cues of what to look for during an assessment.

The primary purpose of this study was to remove the false dichotomy created by traditional binary checklists and present raters with an additional checklist category, which was labelled “attempted” in order to enhance the precision of checklist items. The specific research questions addressed in this study were, how are raters using the “attempted” category when provided with the option in OSCE assessments? What situations are most suited for raters’ use of the “attempted” category? And what impact does scoring of the “attempted” category have on psychometric aspects such as reliability and station pass/fail rates.

**Methods**

To investigate the impact of the “attempted” category, data obtained from an undergraduate OSCE at one medical school were collected and analyzed in two phases. In Phase 1, physicians assessed undergraduate medical candidates in an eight station OSCE using the traditional binary (done/not done) checklist (i.e., business as usual – BAU), the three-point checklist which included “attempted” as an option (i.e., three-point checklist – TPC), or a combination of the traditional binary checklist and the three-point checklist (i.e., business as usual and three-point checklist – BAU/TPC), which required raters to switch from using the traditional binary checklist to the three-point checklist half way through the examination (see...
Figure 4.1). In Phase 2, physicians assigned to a condition that included the use of the additional checklist category were invited to participate in a semi-structured interview regarding their experiences with the additional category.

**Phase 1: OSCE Assessment**

**Participants.** Physicians (n=21) who worked as tutors in the clinical skills component of training participated as the raters in this OSCE. This particular OSCE consisted of seven performance stations and one written station in which candidates’ documented patient information similar to that found in a SOAP (i.e., subjective, objective, assessment, and plan) note. For security reasons, the individual stations cannot be described in more detail; however, the stations can be broadly characterized as a series of milestone appropriate scenarios designed to assess the history taking, physical examination, and technical skills of undergraduate medical candidates (n=102). During each station encounter, medical candidates needed to demonstrate their ability to take a focused history, carry out a physical examination, or perform a particular technical skill.
**Instrument.** Medical candidates and physician raters were randomly assigned to one of three tracks (i.e., BAU, TPC or BAU/TPC). At each station, medical candidates were rated separately by a single faculty rater using an assessment instrument which contained a combination of checklist items and rating scales. By definition, the checklists would be categorized as thoroughness checklists with approximately 20 items per station. The global rating scales consisted of 4 points each and contained specific descriptors that defined each point of the rating scale. These rating scales were designed to capture the overall global performance of medical candidates.

**Procedure.** All physicians who rated in this OSCE attended a rater orientation session. In the rater orientation, the chief examiner explained the purpose of the study and answered any questions. The raters were not given specific instructions on how to interpret the “attempted” category, but they were instructed that the additional category would not negatively affect the assessment outcome of any medical candidates. The raters were also provided with a schedule and their assignments for the OSCE. Once a rater was assigned to a station they remained there for the entire OSCE, which consisted of four rounds of medical candidates. Raters were assigned to one of three tracks: 1) track 1 - BAU - the raters used the traditional binary checklist for all four rounds of students, 2) track 2 - TPC - the raters used the three-point checklist that included “attempted” as an option for all four rounds, or 3) track 3 - BAU/TPC - the raters used the traditional binary checklist for the first two rounds and the three-point checklist for the last two rounds. Regardless of track assignment, all raters used the same 4-point rating scale for the assessment of global performance. Following rater orientation, all raters with the same station assignment met to standardize their approaches to scoring with the content and process of the OSCE.
Phase 2: Raters Written Feedback and Semi-Structured Interviews

Participants. All physician raters (n=21) were asked to complete a feedback form that contained a question specifically about the additional category - *provide a situation where the attempted category was useful or would have been useful as an option*. Raters (n=14) who were assigned to the TPC or BAU/TPC tracks and used the instrument with the “attempted” option were invited to participate in a semi-structured interview following the OSCE examination.

Instrument. Raters who accepted the invitation to be interviewed were asked to describe their approach to rating medical candidates, articulate how they interpreted the meaning of “attempted,” and consider how the “attempted” category influenced their rating behaviour. They were also asked to recall any observations in which the candidate created an emotional reaction for the rater and discuss how they addressed this in the scoring of the candidate. Then they were asked to think about recent and previous experiences rating OSCEs, consider if they ever felt constrained by the scoring system, and comment on the ability of the “attempted” category to free them from those constraints. Finally, the interview concluded with the participants sharing why they choose to participate as raters in OSCEs or other performance assessments.

Procedure. After the OSCE, all raters completed a feedback form. These feedback forms were collected and themed. Although not explicitly instructed, the vast majority of raters, in particular those who used the three-point checklist wrote comments directly on the instrument explaining why they provided a particular evaluation. Of those raters (n=4) who used the “attempted” category and agreed to participate in an interview, each individual rater was sent an electronic list of five questions in preparation for the interview. Each interview took approximately 20-30 minutes and was conducted either in person or over the phone. All interviews were audio recorded to ensure accuracy in reporting.
Results

Phase 1: OSCE Assessment (Quantitative Data)

A total of 52 medical candidates were assessed using the traditional binary checklist and 50 medical candidates were assessed using the three-point checklist. Table 4.1 displays descriptive information for the individual OSCE stations. Item total score correlations were conducted for each station to assess their functioning by checking for inconsistencies across medical candidates’ behaviour. Two items (i.e., one at station 4 for the three-point checklist group and one at station 6 for the binary checklist group) seemed to be flawed and where discarded from the reliability analysis. Reliability measures were calculated for each station for both groups (see Table 4.1). For the majority of the stations, Cronbach’s alpha was higher for the three-point checklists. Reliability values ranged from 0.19 to 0.57 for the binary checklist group and 0.21 to 0.63 for the three-point checklist group; these ranges are comparable to findings from a pilot study (see Smee, Bartman, & Sebok, 2014).

Table 4.1

Descriptive Information and Reliability Indices for OSCE Stations

<table>
<thead>
<tr>
<th>Station Number</th>
<th>Number of Checklist Items</th>
<th>Type of Station</th>
<th>Cronbach’s Alpha for Binary (n=52)</th>
<th>Cronbach’s Alpha for Three-point (n=50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>Physical</td>
<td>0.57</td>
<td>0.63</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>History</td>
<td>0.33</td>
<td>0.46</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>SOAP</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>Physical</td>
<td>0.29</td>
<td>0.21*</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>History</td>
<td>0.45</td>
<td>0.34</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>Technical Skills</td>
<td>0.19*</td>
<td>0.41</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>Physical</td>
<td>0.47</td>
<td>0.57</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>History</td>
<td>0.38</td>
<td>0.48</td>
</tr>
</tbody>
</table>

*Indicates that one checklist items was removed because the item was producing a negative correlation with the overall the item total score.
Characteristics for the binary checklist and three-point checklist groups are shown in Table 4.2. The mean for the binary group was 80.80, while the mean for the three-point group was 77.14. To test whether these means differed, an independent sample t-test was conducted. The mean difference was statistically significant \( t(100) = 4.77, p < .001, d = 0.94 \).

Acknowledging that this comparison included two separate groups of raters, a t-test was also conducted within a single track that contained the same raters. The BAU/TPC track in which raters used both checklist instruments had a mean score of 79.84 for the binary group and 76.93 for the three-point group. The difference between the binary and three-point checklist groups was not significant \( t(32) = 1.92, p = .063, d = 0.66 \). It is worthwhile to note that although this mean difference was not statistically significant, its absolute size was similar to that found for the full comparison, suggesting the lack of significance may have been due to insufficient power \( (n=18 \text{ and } n=16) \).

Table 4.2

<table>
<thead>
<tr>
<th>Rater Group</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>t-stat</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Raters, All Tracks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binary</td>
<td>52</td>
<td>80.80</td>
<td>3.65</td>
<td>4.77</td>
<td>100</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Three-Point</td>
<td>50</td>
<td>77.14</td>
<td>4.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined BAU/TPC Track</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binary</td>
<td>18</td>
<td>79.84</td>
<td>4.03</td>
<td>1.92</td>
<td>32</td>
<td>.063</td>
</tr>
<tr>
<td>Three-Point</td>
<td>16</td>
<td>76.93</td>
<td>4.81</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 provides the descriptive statistics for each station depending on how the additional checklist option of “attempted” was scored. Means and standard deviations were calculated by adopting an institutionally traditional weighting of 80% for the checklist and 20% for the global rating. The means and standard deviations were comparable within stations, but varied across stations. Also shown in Table 4.3 is the number of candidates who would have failed the station.
when the station cut score was set to 60%. Depending on how the “attempted” category was scored, the difference in number of failures within a station ranged from 0 to 34.

Table 4.3

Descriptive Measures for Differences in the Scoring of “Attempted” among Candidates

<table>
<thead>
<tr>
<th>OSCE Station</th>
<th>Binary Checklist (Control Group)</th>
<th>“Attempted” Scored as 0.0</th>
<th>“Attempted” Scored as 0.5</th>
<th>“Attempted” Scored as 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Station Mean (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 – Physical Station</td>
<td>94.76</td>
<td>94.09</td>
<td>95.53</td>
<td>96.97</td>
</tr>
<tr>
<td></td>
<td>8.32</td>
<td>8.79</td>
<td>6.46</td>
<td>4.75</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 – History Station</td>
<td>84.67</td>
<td>81.33</td>
<td>81.37</td>
<td>81.40</td>
</tr>
<tr>
<td></td>
<td>7.87</td>
<td>9.99</td>
<td>10.03</td>
<td>10.08</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4 – Physical Station</td>
<td>84.83</td>
<td>75.43</td>
<td>78.89</td>
<td>82.36</td>
</tr>
<tr>
<td></td>
<td>8.78</td>
<td>8.30</td>
<td>6.93</td>
<td>6.73</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5 – History Station</td>
<td>81.21</td>
<td>79.41</td>
<td>81.65</td>
<td>83.89</td>
</tr>
<tr>
<td></td>
<td>11.36</td>
<td>11.52</td>
<td>10.08</td>
<td>10.08</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6 – Technical Skills Station</td>
<td>77.83</td>
<td>77.86</td>
<td>81.42</td>
<td>84.99</td>
</tr>
<tr>
<td></td>
<td>10.91</td>
<td>13.29</td>
<td>10.98</td>
<td>9.49</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>7 – Physical Station</td>
<td>82.29</td>
<td>79.20</td>
<td>81.96</td>
<td>84.71</td>
</tr>
<tr>
<td></td>
<td>9.86</td>
<td>12.22</td>
<td>10.70</td>
<td>10.37</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8 – History Station</td>
<td>61.26</td>
<td>55.49</td>
<td>57.86</td>
<td>60.23</td>
</tr>
<tr>
<td></td>
<td>9.10</td>
<td>11.30</td>
<td>10.50</td>
<td>9.94</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>34</td>
<td>29</td>
<td>22</td>
</tr>
</tbody>
</table>

\( a = 102 \) Candidates
Phase 2: Raters Written Feedback and Semi-Structured Interviews (Qualitative Data)

From the raters’ feedback forms and comments written on the checklist instruments, there appeared to be three themes that encompass how raters interpreted the “attempted” category. These themes were: 1) something unusual happened, 2) verbalized, but did not do, and 3) attempted a task, but not done correctly. Additionally, one rater also wrote on the feedback form that “attempted” could be used in situations where a candidate was in the process of correctly doing a task, but was unable to complete the task because they ran out of time.

**Something unusual happened.** Raters described feeling challenged when something outside of the ordinary occurred. Assessment instruments are rarely designed to accommodate the unanticipated behaviours of medical candidates. Within the OSCE setting, raters identified unusual events that occurred while candidates were taking a history, conducting a physical examination, or performing a technical skill. For example, while taking a focused history, several raters described an instance where medical candidates correctly discussed confidentiality, but did so at the middle or end of the history. For physical examination stations, raters commented on candidates who did unusual things such as using the thumb to palpate. One rater checked “attempted” for a particular task and wrote beside the checklist item “one handed technique, unusual.”

**Verbalized, but did not do.** Throughout the OSCE assessment some medical candidates would verbalize what they intended to do in order to help cue the rater that they were about to perform or had performed a particular task. Raters identified examples such as medical candidates who verbalized that they were going to inspect the neck, but did not actually perform an inspection. Likewise in history taking, medical candidates would verbalize their mental checklist of questions, but failed to ask for such information from the standardized patient.
Accompanying the use of “attempted” on the checklist instrument for these situations were comments such as “verbalized, did not do” and “only verbalized; did not really inspect.”

**Attempted a task, but not done correctly.** Perhaps the most intuitive use of the “attempted” category emerged in situations where something was done partially or insufficiently. Raters described instances in which medical candidates’ suture technique were generally good, but were a little loose or where medical candidates were expected to ask about three associated symptoms, but only asked about one or two. Sometimes the technique was correct, but the medical candidate was feeling or palpating the wrong area. The most interesting aspect of this theme was the extent to which variability occurred. With the first two themes, raters typically evaluated these instances as “not done” or “attempted.” However, when medical candidates attempted a task, but did not complete it satisfactorily raters used all three checklist options to communicate a similar situation. For example, some raters checked “not done” and said “discussed half way through the interview” or “mentioned confidentiality late and didn’t clarify exception.” Meanwhile, other raters checked “attempted” and commented “don’t forget exceptions to confidentiality” or “exceptions to confidentiality, worried for your safety or someone else.” Other raters checked “done” and wrote “at end of interview” or “did not mention need to break confidentiality if risk identified.” It is important to note that the two raters who checked “done” were on the BAU track, which means they did not have “attempted” as an option.

The follow-up interviews with select raters echoed the findings from all of the raters’ written feedback and allowed for a more specific, in-depth examination of the “attempted” category. During the interviews, raters commented on how they interpreted and used the
“attempted” category, identified situations most suited for the “attempted” option, and discussed the impact of three-point checklists.

**Raters’ interpretations and use of the “attempted” category.** Not surprisingly, raters commonly used the “attempted” category as a medial between done and not done. Most frequently, raters commented on how “attempted” was useful when a medical candidate was on the right track, but some element was missing:

“I think that the students had the right approach, or some of the right elements, but didn’t have the totally appropriate approach or all the elements of the exam in place” – Rater 1, Station 7

“If you say attempted that means you did this part well, but you need to work on this” – Rater 2, Station 4

“The attempted, I was using it as sort of an in between mark. Something that is done partially, but not completely” – Rater 3, Station 6

“The instep between not doing it and doing it and how I would have judged it as intermediate was a gut feeling” – Rater 4, Station 8

Raters also spoke about “attempted” as an opportunity to give students part marks. Many raters wanted to acknowledge and give credit when a candidate did something well, but may have missed a few things. Additionally, raters spoke about “attempted” as a way of providing part marks and feedback to a candidate when an aspect was done, but needed further attention. Often these instances were determined through a “gut feeling” or an urge to “jump in” during the course of the assessment.

**Situations most suited for the “attempted” option.** When asked about which situations are most appropriate for using the “attempted” option, the raters identified several different situations. One rater spoke about using “attempted” in formative OSCEs so that medical candidates can be given feedback to improve their skills:

“Formative OSCEs because sometimes it takes a little tweak and the student gets it” – Rater 1, Station 7
Meanwhile, another rater suggested that “attempted” could be used when candidates complete all checklist items, but lacked the humanistic element in their approach:

“They may have gotten all the checklist items, but the way they did the exam was pretty shoddy, the approach was by rote” – Rater 2, Station 4

Some raters focused on situations where there is a lack of consistency between what candidates say they are doing in an OSCE station and what they are actually doing. Consistent with the comments observed in raters’ written feedback, raters struggle to make evaluative judgments in situations where a candidate’s action contradicted their verbalized intent:

“I would use attempted, if they verbalized a different intent” – Rater 3, Station 6

Finally, as previously identified in the literature (see McKinley et al., 2008), raters struggle with binary checklists especially when they were unclear or complex. Complex checklist items are those that contain “and” statements. Based on the principles of logic, an “and” statement requires both conditions to be true in order for the statement itself to be true. Within the context of OSCE assessments complex binary checklist items are particularly problematic for raters because they struggle in situations where candidates only completed half of the task:

“I think in most of the items, the done or not done is fairly clear, so the attempted category shouldn’t have impacted the scoring decisions; it might in items where the criteria was less clear, I’d be more inclined to write attempted rather than successful” – Rater 4, Station 8

The impact of binary and three-point checklists on raters’ behaviours. Raters offered several additional comments that highlighted their personal feelings and experiences with the use of the “attempted” category. Many raters spoke favorably about the “attempted” category, especially given the struggle raters experience when something was partially completed and the only available options were “done” and “not done.” Several raters felt that false dichotomies
(e.g., pass/fail or done/not done) restrained them in terms of articulating what genuinely occurred within the context of an assessment:

“Pass/fail doesn’t capture the experience and so yeah I feel constrained by the scoring system” – Rater 1, Station 7

“I found it very helpful because you were able to give them part marks if you will, whereas if it’s all or nothing you’re often struggling” – Rater 2, Station 4

“It’s nice to have the attempted category, it decreases the cognitive load for me because having to make the decision between done or not done in someone who is partially done is difficult” – Rater 4, Station 8

Finally, raters described the “attempted” category as being more accurate and representing a fair assessment of medical candidates. Throughout the interviews, as well as during the OSCE, many raters commented on the subjectivity in performance assessments and how they try to be more objective and fair in their evaluations of medical candidates:

“I thought it [attempted] was fair to the students and more accurate” – Rater 2, Station 4

“I think it does help because again it gives that halfway mark and it gives a chance to say they did it, but not to a fully competent level, they’re on their way there. It gives me that opportunity to give them a point and gives them an opportunity to score better because I do think it’s unfair to say yes or no for many of those evaluation points” – Rater 3, Station 6

“I’d say overall the attempted is a more positive addition because I think it’s fair to say to students you partially completed” – Rater 3, Station 6

**Discussion**

Within medical education it has been understood for some time that medical competence is easy to recognize and hard to articulate for the purpose of setting measurement standards. Currently, many of the assessments used to evaluate clinical skills are inadequate in their ability to obtain precise and meaningful information from raters about medical candidates. Many raters struggle to translate the degree to which a performance is completed, especially in situations where a candidate carried out a task in a mediocre manner. When forced to use traditional binary
checklists, raters describe struggling as to how to best reflect performances of those medical candidates in the “grey zone.” This study explored the potential of an additional “attempted” checklist category in an effort to enhance raters’ abilities to articulate a trustworthy basis for their evaluative judgments.

Psychometrically, when the false dichotomy created by traditional binary checklists was removed, we found that raters were able to further discriminate among the performances of medical candidates, which in turn provides greater precision in raters’ judgments during OSCE assessments. The “attempted” category appeared to reduce the frequency of raters using the “done” category as a default for candidates who only partially completed a task.

A critical finding from this study was how much variability in OSCE ratings depends upon how the “attempted” category was weighted and scored. In some cases, the scoring of “attempted” varied by less than 1%; however, in other instances, depending on how “attempted” was scored, candidates’ scores could vary by almost 10%. Additionally, the number of candidates who would have failed a particular station ranged from 0 to 12. This suggests that the “attempted” category is capturing worthwhile information about borderline candidates.

In terms of accuracy and ease of use, raters’ written feedback and post-OSCE interviews revealed that removing the false dichotomy made OSCE assessment easier, especially in cases where candidates did something unusual, were inconsistent in their performance, or only partially completed a task. The raters described feeling like the “attempted” category better reflected what happened during the OSCE station, and thus enabled them to provide a more accurate assessment of medical candidates’ performance.

On a final note, the raters who participated in this research noted that the attempted category could be used as a means of providing feedback (through partial credit) to candidates
who were on the right track, but had missed some element. Previous research indicates that some raters refuse to fail candidates because they are not confident in their assessment abilities or they do not want to justify and defend poor or negative evaluations (e.g., Dudek et al., 2005). However, raters’ need and desire to award partial marks seemed to stem from raters not agreeing with the fundamental purpose and use of OSCE assessments. The majority of the raters in this study expressed how important it was for the OSCE assessments to be fair for the candidates. Many raters worried that their own subjectivity might negatively impact a candidate, and thus, followed the checklists closely in order to ensure their evaluations were fair. However, if raters are not engaged with the assessment process from the beginning or the checklist items are unclear, too complex, or ill-defined, then assessing with binary checklist creates unwanted error. Having the “attempted” category gave raters an out, it allowed them an opportunity to better identify the difficult aspects of assessing medical competence.

Limitations

Each medical candidate was assessed by only one rater per station and there was no overlap between the candidates or the raters. Although medical candidates and raters were randomly assigned, there were no previous measures to ensure that the groups of candidates were equal to each other prior to the OSCE. Additionally, the checklists used in this particular study were developed for this exam specifically and had not previously validated. Finally, the small sample size for the comparison of candidates on the BAU/TPC track likely resulted in a loss of power, which led to no statistically significant differences observed between the two groups of candidates evaluated by the same group of raters. Furthermore, given the few number of raters interviewed post-OSCE it is possible that additional interpretations of the “attempted” category may also exist.
Conclusions

To summarize, using a three-point checklist, which included an “attempted” category, allowed raters an opportunity to more accurately identify and articulate what occurred in situations when it is difficult to make a definitive evaluation of medical competence. Findings from this study illustrate that the use of an additional checklist category can have an impact on psychometric factors such as pass rates. Overall, raters felt that the “attempted” option supported a more authentic and fair assessment of candidates’ medical competence.
References


Abstract

Purpose: The shift towards broader, programmatic assessment has revolutionized the approaches taken by many in the assessment of competence. In order to understand the association between quantitative and qualitative evaluations, this study aims to explore the relationships that exist between assessors’ checklist scores, ratings, and written comments.

Methods: Data from the McMaster Modular Assessment Program (McMAP) were collected and analyzed using regression analyses. Additionally, narrative comments corresponding with the “done” and “done, but needs attention” checklist categories were further analyzed using content analysis.

Results: The regression analyses revealed that the task rating, provided by the faculty assessor, significantly predicts the use of the “done, but needs attention” checklist category. Analyses also identified that the “done, but needs attention” category was a significant predictor in determining whether a narrative comment will be balanced. Finally, a close analysis of qualitative comments revealed differences in the comments provided to residents.

Conclusions: This study highlights some of the relationships that exist between checklists, rating scales and written comments. Our findings highlight that task ratings and global ratings are differentially associated with checklist options. Furthermore, analysis of written comments supports the notion of a “hidden code” used to communicate assessors’ evaluation of medical competence. This study has implications for how individuals should be interpreting information obtained from qualitative assessments.
Mixed Messages or Miscommunication? Investigating the Relationship between Assessors' Workplace-Based Assessment Scores and Written Comments

The assessment of competence at various levels of medical educational training continues to be an area of growing interest. The shift towards broader, programmatic assessment has revolutionized the approaches taken by many in the evaluation of competence. Programmatic assessment acknowledges the limits of human judgement and combines information from various assessment measures in order to enhance the quality of competency assessments (Schuwirth & van der Vleuten, 2011). According to Schuwirth and van der Vleuten (2012), assessment of medical competence should be three things: fair, transparent, and defensible. As many medical schools and external organizations strive to develop assessment programs that meet the aforementioned criteria, there exists an ongoing need to better understand the relationships that exist between various quantitative and qualitative indicators of competence.

Previous research has identified some of the pitfalls of relying solely on a quantitative psychometric framework for competency evaluation (Ginsburg, McIlroy, Oulanova, Eva, & Regehr, 2010; Govaerts & van der Vleuten, 2013). As studies continue to highlight shortcomings of quantitative approaches for the assessment of clinical competence, there exists a greater push towards the inclusion or use of qualitative data such as performance reports and narrative comments. Existing literature examining written comments indicates that narratives: 1) correlate with assigned grades (Plymale, Donnelly, Lawton, Pulito, & Mentzer, 2002), 2) can be used to assess learning and rank residents (Ginsburg, Eva, & Regehr, 2013), and 3) provide a more holistic impression of an attending’s competence (Ginsburg, Gold, Cavalcanti, Kurabi, &McDonald-Blumer, 2011).

While there has been a great deal of support for the use of narrative comments as an evaluation component in assessment programs (Ginsburg, Regehr, Lingard, & Eva, 2015),
qualitative measures are not without challenges. Several studies have highlighted the issue of quality with written comments (i.e., the quality is typically not high; Bismil, Dudek, & Wood, 2014; Canavan, Holtman, Richmond, & Katsufrakis, 2010; Dudek et al., 2012). Furthermore, some individuals are reluctant to provide written comments (Canavan et al., 2010), especially when the information is constructive and shared with the assesse, often because they anticipate that their comments will be challenged or appealed (Dudek, Marks, & Regehr, 2005).

In addition to issues related to the quality of narrative comments, there also exists another predicament. Within the context of medical education, many individuals conflate qualitative assessments (e.g., written comments) with feedback, which is problematic because assessment and feedback can have very different purposes. In the context of medical education, assessment is typically viewed as a summative activity used to report on the learning that has been achieved; not to mention, it can also serve as an accountability aspect for various stakeholder groups. Meanwhile, feedback is seen as formative, a crucial aspect of teaching required to support learners throughout future levels of training. As a result of this integration, and in an effort to satisfy both the assessment requirements and the desire to provide valuable feedback to learners (van der Leeuw, Overeem, Arah, Heineman, & Lombarts, 2013), assessors have developed methods of communicating using a “hidden code” (Ginsburg et al., 2015). This hidden code is currently being used by assessors to satisfy the reporting and accountability requirements without negatively impacting learners. This is achieved by purposeful omission of elements and vague use of language. For example, if an assessor provided the comment, “Good resident, improving well” that would be an important signal in negatively interpreting the assessment of this resident. Comparatively, narrative comments such as “I would hire him” and “Excellent shift – saw many complex patients. Ready to be a senior resident. Managed a difficult airway efficiently and
impressively” would signal a positive interpretation of a resident’s ability. The difficulty with the use of a hidden code is that those unfamiliar with the language and context may overlook or misinterpret assessors’ intended meanings. Furthermore, this discourse can also make it appear as though the information obtained from quantitative and qualitative assessments are misaligned.

In order to capture and report more accurate information about individuals, assessment programs need to find a way to align quantitative and qualitative evaluations. Current research investigating the association between quantitative and qualitative measures of medical competence is essentially nonexistent. One explanation for this gap in the literature relates to the discourse surrounding competency-based education, where a false dichotomy of quality over quantity has been created (ten Cate, 2015). In an attempt to close this gap, this study aims to explore the relationships that exist between assessors’ quantitative and qualitative evaluations of medical competence. The following research questions were addressed by this study: 1) What is the relationship between checklists, rating scales, and written comments? 2) How well do checklist items and global ratings predict the quality of narrative feedback that could be used to provide meaningfully relevant information?

Method

Participants

Data for the current study were obtained from residents in their first or second year of postgraduate training. These residents were part of McMaster University’s Emergency Medicine postgraduate program and were assessed using the McMaster Modular Assessment Program (McMAP; Chan & Sherbino, 2015). McMAP is comprised of various workplace-based assessments (WBAs) blueprinted to the CanMEDS framework. During postgraduate clinical training, WBAs specifically developed for each of the CanMEDS roles, were completed by
faculty members in order to assess the readiness of postgraduates for unsupervised medical practice. In total, there were seven postgraduate year one (PGY1) tasks and seven postgraduate year two (PGY2). The resulting sample consisted of 342 McMAP assessments. Each assessment task contained a checklist, performance ratings, and written comments.

**Instruments and Procedure**

As previously mentioned, McMAP is comprised of various WBA tools blueprinted to the seven CanMEDS roles (i.e., medical expert, communicator, collaborator, professional, scholar, manager, and health advocate). Each assessment task consists of five parts: 1) A task checklist with four checklist options (i.e., done, done but needs attention, not done, and N/A for case); 2) A task rating instrument where a rating is provided for the specific task; 3) A task comment where the assessor is required to provide narrative comments to support the task rating and offer next steps before proceeding to the next section of the assessment task; 4) A global rating designed to capture a resident’s holistic performance throughout a shift; and finally 5) A global comment where the assessor can comment on a resident’s overall performance in order to prompt formative feedback. For an example of a complete McMAP assessment task, please see Appendix 1 (Chan & Sherbino, 2015). At particular milestones throughout residents’ postgraduate training, these WBAs are compiled and evaluated in order to assess and monitor the readiness of postgraduates for unsupervised practice.

This study only used McMAP assessment tasks from the communicator and collaborator roles. All of the written comments from 342 McMAP assessments were also assessed by an expert specialized in both education and Emergency Medicine content. Using the completed clinical evaluation report rating (CCERR; Dudek, Marks, Wood, & Lee, 2008), the expert
evaluated both the task and global comments using a 5-point rating scale. This rating scale was anchored at three points (i.e., 1 = “Not at all”, 3 = “Acceptable”, 5 = “Exemplary”).

**Regression Analyses (Quantitative)**

In order to examine the relationship between checklist, rating scales, and narrative comments, logistic regression analyses were conducted on data from the 342 McMAP assessments using R, version 3.1. The main logistic regression model was constructed to determine which components predicted if the “done, but needs attention” checklist option was used. This model explored the following predictors: task rating and global rating (as assessed by the faculty member during residency training), task strength and task weakness ratings (as assessed by the established expert), and the task written comment (as assessed by the length of comment).

To validate the results of the first model, an ordinal logistic regression was built to further explore the relationship between done, but needs attention and residents’ overall strengths and weaknesses as assessed in the global comments. Using the established expert’s ratings on the CCERR, comments were evaluated according to the degree in which they balanced strengths and opportunities for improvement. The two predictors included in this model were “done, but needs attention” and the faculty assessor’s global rating of the resident. Both regression models were validated to test the underlying model assumptions prior to interpretation of results.

**Content Analysis (Qualitative)**

Subsequent to the aforementioned regression analysis, all task and global comments were collected, reviewed, and assigned to one of two categories: “done” or “done, but needs attention.” These two categories of comments were then closely analyzed and coded using the positive and negative red flags identified by Ginsburg and colleagues (2015) to better understand
the types of comments that typically accompany situations in which a resident fully completed a task or completed a task but still required attention. This portion of the analyses was carried out by two of the researchers.

**Results**

A total of 158 PGY1 and 184 PGY2 McMAP assessments, designed to evaluate communicator and collaborator CanMEDS roles of residents, were analyzed using logistic regression and content analysis. Of the 342 McMAP evaluations, 104 (30.4%) of them were missing a task comment and 93 (27.2%) were missing a global comment. This finding of missing comments is interesting because in the system’s current electronic format, assessors are required to provide narrative comments (to expand upon their numerical scores) before they can proceed to the next portion of the assessment. Nevertheless, assessors found a way to avoid providing comments when they felt it was appropriate.

**Quantitative Findings**

After examining the assumptions of logistic regression, a direct logistic regression was performed in R to analyze the outcome (i.e., if “done, but needs attention” was used in the checklist portion of the McMAP assessment). Both logistic regression models were constructed with data from all 342 McMAP assessments. In cases in which a task or global rating was missing, the entire assessment was removed from the analysis. A test of the logistic regression model examining the use of the “done, but needs attention” checklist option was statistically significant, $R^2 = 0.306$, $\chi^2(5, N = 321) = 68.27$, $p < .001$. This suggests that some of the five predictors were able to successfully predict when the “done, but needs attention” checklist option would be used. Table 5.1 shows the regression values for each individual predictor.
According to the Wald criterion, our expert’s ratings of strengths and weaknesses within individual comments, using the CCERR, appeared to be reliable predictors of when the “done, but needs attention” checklist option was used (see Table 5.1). Additionally, the most reliable predictor of “done, but needs attention” usage was the task rating of the resident provided by the faculty assessor at the time of assessment. The length of the task comment and faculty assessors’ global ratings were not statistically significant predictors according to this model. These results support the interpretation that when faculty assessors noted and commented about weaknesses in a resident’s performance, they were also more likely to use the “done, but needs attention” checklist category. Likewise, when faculty assessors provided more specific strengths in their comments, they were less likely to use the “done, but needs attention” category. Finally, a relationship was observed between faculty assessors’ task rating and their use of the “done, but needs attention” checklist option (i.e., the higher the task rating, the less this category was used).

To substantiate the results of the first model, a second ordinal logistic regression model was analyzed to examine the relationship between the “done, but needs attention” checklist category, residents’ global rating scores, and global written comments. The dependent measure used in this analysis was the expert’s rating (using the CCERR) of balance within global comments, more specifically the extent to which global comments provided both strengths and

### Table 5.1

*Regression Predictors for Checklist Option, “Done, but needs attention”*

<table>
<thead>
<tr>
<th>Model Predictors</th>
<th>Beta</th>
<th>S.E.</th>
<th>Wald</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.95</td>
<td>1.32</td>
<td>3.75</td>
<td>0.0002</td>
</tr>
<tr>
<td>Task Weaknesses Rating</td>
<td>0.36</td>
<td>0.14</td>
<td>2.42</td>
<td>0.016</td>
</tr>
<tr>
<td>Task Strengths Rating</td>
<td>-0.41</td>
<td>0.12</td>
<td>-3.47</td>
<td>0.0005</td>
</tr>
<tr>
<td>Task Rating (Faculty)</td>
<td>-1.06</td>
<td>0.29</td>
<td>-3.61</td>
<td>0.0003</td>
</tr>
<tr>
<td>Task Comment Length</td>
<td>0.013</td>
<td>0.031</td>
<td>0.40</td>
<td>0.69</td>
</tr>
<tr>
<td>Global Rating (Faculty)</td>
<td>0.026</td>
<td>0.27</td>
<td>0.09</td>
<td>0.92</td>
</tr>
</tbody>
</table>
areas of improvement. This model was statistically significant, $R^2 = 0.041$, $\chi^2 (2, N = 321) = 10.67$, $p = .005$. However, the only significant predictor in this model was the “done, but needs attention” checklist option ($b = 0.045; p = .001$), which suggest that the more “done, but needs attention” was used, the higher the amount of balance found within the comment. This finding indicates that balanced comments tended to occur in situations in which “done, but needs attention” was used. This finding suggests that the “done, but needs attention” category may be encouraging faculty to defend or explain their judgments by elaborating on what was done well (strengths) and what requires further care and consideration (areas of improvement), thus justifying the category score.

**Qualitative Findings**

The analysis of task and global McMAP comments supported the claim that assessors communicate using a hidden code that requires reading between the lines (Ginsburg et al., 2015). Written comments associated with the “done, but needs attention” checklist option read quite different than those accompanying the “done” category. Narrative comments in the “done, but needs attention” transcripts often focused on the task, even if they were intended to be global comments. The word, “good” was used most frequently in the “done, but needs attention” comments and in most instances it was followed up with an explanation that noted a deficit.

- “Time, name, good assessment note. [Reassessment] notes need to be expanded and documented.”
- “Good knowledge. Hardworking, good management plans. Work on charting legible notes.”
- “Organized and good plans. Still need to focus on some presentations.”
- “Overall good shift. Saw lots of patients, good with [Ultrasound]. Continue to improve your clinical exposure by seeing as many patients as possible and accepting feedback.”
It appears the “done, but needs attention” option helped to establish a framework (i.e., a statement of strength and an area of improvement) that assessors used when providing written comments. Words commonly appearing in the “done, but needs attention” comments were: needs, consider, more, remember, follow-up, good, and hardworking/works hard.

In contrast, written comments associated with the “done” checklist option carried a different connotation than the “done, but needs attention” category. These comments were also longer in length. The comments corresponding with “done” tended to focus mainly on positive aspects that reinforced the checklist option, used writing conventions (punctuation marks or capital letters) to provide emphasis, and tended to focus on the resident in relation to others, suggesting more norm than criterion referencing.

“On own initiative made several attempts to contact family of unstable patient. Good foresight!”

“Excellent shift – saw many complex patients. Ready to be a senior resident. Managed a difficult airway efficiently and impressively. Also managed a septic patient very well. Inserted a central line extremely well and appropriately managed a patient with multiple system failure. Impressive work!”

“Excellent consult skills. Organized, thorough, and reliable. Spoke with multiple consult and obtained approach clinically and relevant advice. EXCELLENT.”

“Beyond her level of functioning in the department. Excellent charting/assessments and reassessments. Great management/discharge/plans. No concerns. Great job! Very efficient for her level – trying to instruct medical student as well.”

“Functions extremely independently and well for PGY1. Probably more like a PGY2. Concise histories, good plans and treatment and disposition plans.”

Written comments that elaborated on the “done” category offered some observation or assessment, but the suggestions were stated more passively than those associated with “done, but
needs attention.” They also often made reference to some global aspect of the resident’s performance in a way that could help the resident achieve a higher level.

“Suggested could improve on assertiveness and confidence as he is a bit quiet.”

“Good job with DKA presentation. Feel free to provide clinical significance for lab findings.”

“Proficient. Efficient. Reliable. Disposition plans earlier in management if possible.”

“Clear explanations. Ensured plan understood. Offered opportunity for questions. In future, give results of test/investigations before transitioning into discharge plans.”

Within the comments corresponding with “done,” many assessors appeared to focus on the resident’s current level of performance and the “next steps” a resident could achieve to move onward. In several instances, faculty assessors would indicate the resident was performing well by stating “no concerns” in the comment fields.

**Discussion**

The purpose of this paper was to examine the relationships that exist between checklists, rating scales, and written comments in order to better understand how these assessments could be used to assess learning and provide feedback. McMAP assessments function as a programmatic approach for evaluating the medical competence in the workplace. McMAP assessments contain both score data and narrative comments, which optimize the way in which assessors can provide information about residents. Results from this study indicate that even when assessors were required to provide written comments, many of them (approximately 30%) found a way to avoid giving such information. Despite the opportunities provided, and the wide acceptance and appreciation of narrative feedback, some assessors were still unwilling to supplement their scores with written comments. For a nearly a decade, we have known some the reasons why some
assessors are not willing to provide narrative feedback, especially in situations where it may be negative (Dudek et al., 2005); however, despite our best efforts to make assessments more accessible, there is still resistance from some assessors.

Results from the logistic regression analyses further highlight the relationship between checklist scores, rating scales, and narrative comments. In the main logistic regression model it is interesting that the task rating was significant, but the global rating was not. This may have occurred because the task rating was capturing task or content specificity that is not always detected with a global rating score. Within the medical education literature, it is well-known that task sampling impacts the reliability of scores as well as increases variability (e.g., Swanson, Clauser, & Case, 1999); therefore, by allowing assessors the opportunity to provide both a task rating and a global rating, more specific information related to the resident’s competence for a particular task can be obtained. Given that both the logistic regression models appeared to show that the “done, but needs attention” checklist option was associated with measures of balanced comments and formative feedback, it is possible that the “done, but needs attention” category is identifying residents with “emerging competence” who exhibit distinguishable strengths and weaknesses throughout a single encounter or shift. From a decision-making perspective, the “done, but needs attention” checklist option seemed to serve as a platform to solidify assessors decisions by encouraging them to provide narrative comments about a resident’s competence and abilities to justify their assessment (Sibbald, de Bruin, Yu, & van Merrienboer, 2015).

A close analysis of the data collected from McMAP assessments by two individuals’ revealed substantive differences in the type of comments and feedback provided in WBAs. General feedback that challenges a resident to move forward more holistically was more commonly associated with higher checklist scores and ratings. Meanwhile, feedback that was
task or context specific was aligned with moderate scores and the checklist option, “done, but needs attention.” Overall, the results of this study support the findings of Ginsburg and colleagues (2015) who suggested the presence of a hidden code when reading and interpreting narrative comments. These findings highlight an assessment challenge (i.e., one of communication and translation) that needs to be further explored in order to examine how “shared” this understanding of the hidden code really is among individuals in the medical education community.
References


Imagine for a moment that you are watching a hockey game, in addition to the millions of fans watching, there are also linesmen, referees, and expert officials that are responsible for ensuring that the game is fair, the transparent rules are enforced, and the decisions are defensible. All of these individuals are watching the same game; however, each person has a different role that requires them to pay attention to specific details. The criteria for evaluating whether an individual is best suited to be a referee or linesman can be determined by the level of expertise and experience a particular individual possesses. An individual can be a linesman in the playoffs and a referee in the regular season because context and stakes matter. Bringing this example back to raters in the context of performance assessments, this dissertation focused on raters’ behaviours in order to determine how the information they provide could be used more purposefully.

Given the shift towards competency-based assessment and programmatic assessment, the three studies described in this dissertation examined rater variability by attempting to understand what raters are doing as they make evaluative judgments at various medical education milestones. Admissions (i.e., entry into medical school), undergraduate (i.e., core medical education training), and postgraduate (i.e., specialized medical education training) contexts represent three crucial points within medical education. Each study examined a different aspect of raters’ interaction with various assessment instruments within the context of medical education. The foundation for examining the manner in which raters communicate and articulate themselves is warranted given the impact that raters’ judgments and subsequent interpretations can have on decision-making. Schuwirth and van der Vleuten (2012) maintain that assessments should be fair, transparent, and defensible. Fairness refers to assessments that are free from bias.
and do not directly or indirectly advantage a particular group or individual over another. Fairness could also encompass assessments that are suitably matched to the purpose for which they are designed. Transparent assessments are those that are easy to understand, openly described, and contain no surprises. Finally, defensible assessments are those that are justifiable and can be easily defended using valid and concrete evidence.

Challenging these fundamental principles of assessment are critical aspects such as context, purpose, and complexity. Berendonk, Stalmeijer, and Schuwirth (2013) showed that raters’ perceptions of the assessment context, task, or purpose can greatly impact the quantity and quality of information obtained from raters. Therefore, it is important to examine how to obtain accurate information from raters while also considering the role various aspects such as context can have on raters’ judgments and subsequent decision-making. Many judgments and decisions are based on future consequences, which could adversely alter raters’ interpretations during the course of an assessment. Without the ability to collect information from raters that is without influence from external factors or motivated by perceived consequences, evaluative judgments are of limited value in supporting assessments that are fair, transparent, and defensible.

**Summary of Findings**

The research presented in the three preceding chapters begins to examine how raters are seeing things and what they are doing in different assessment environments. The first study (Chapter 3) - *Seeing Things Differently or Seeing Different Things? Exploring Raters’ Associations of Non-Cognitive Attributes* - examined the underlying behaviours of raters as they assessed the non-cognitive competencies of prospective medical students using the MMI. At some stations (e.g., Station 6) there was little to no variability between raters. At other stations
(e.g., Station 4), raters varied significantly from each other in terms of their scoring behaviours. Closer analysis of this phenomenon revealed that in situations where there was low variability, all of the raters were making the same associations among the non-cognitive attributes. For example, every rater would see communication, critical thinking, and empathy as three distinct attributes (see also Figure 3.2). In situations, where there was high variability, one or more raters had differentially associated the non-cognitive attributes (see also Figure 3.3), suggesting that they were seeing different things by viewing one attribute more distinctly than the others. It also occurred that some raters were unable to discriminate and distinguish among the different attributes.

The combination of the MFRM and hierarchical clustering offered an opportunity to better understand individuals’ rating behaviours, which could be used to further examine idiosyncratic differences in raters by establishing when inconsistencies are a result of paying attention to different things or having different interpretations of the same thing. Findings from this study also suggest that some raters are better at distinguishing among different non-cognitive competencies than others given the restraints of our current assessment system. Depending on the established purpose, some raters may be well-suited for holistic rating, while others could be more effective when there is a need to capture nuanced differences in performance. This study provides a foundational mechanism that can be used to help explain rater variability by identifying when raters are seeing the same thing differently and when they are simply seeing different things. While much of the literature up to this point has debated whether rater variance is a good thing or a bad thing within assessment (Gingerich, Regehr, & Eva, 2011), this study focused on identifying, with greater precision, a source of rater variance and then asking under what conditions such variance would be useful. For instance, when it comes to giving feedback,
variability in raters’ perspectives could lead to a more comprehensive account of how a learner can improve, especially if the variability is derived from seeing different things.

A more in-depth examination of how raters may see things differently was reported in the second study (Chapter 4) - *Understanding the Shades of Grey between “Done” and “Not Done” in Objective Structured Clinical Examination (OSCE) Assessments*. In this study, a third checklist option (i.e., attempted) was added to existing assessment instruments in order to create a three-point checklist that raters could use to assess the degree to which a task was completed. The “attempted” category provided an alternative to the falsely created binary of “done” and “not done” and, in doing so, eliminated the need for raters to force a response into a category of ‘best fit’ where in fact no true fit may have existed. Quite conceivably, this “attempted” category resolved some raters’ cognitive dissonance as they were able to note inconsistencies in a trainee’s performance and use such information to justify their decision about the trainee’s overall level of competence. Differences were found between raters that used the traditional binary checklist and raters that used the three-point checklist (p < .001), with the binary checklist group exhibiting a higher mean and slightly more variability. The extent to which these differences impacted the scores and pass rates of medical candidates varied depending upon how the “attempted” category was weighted and scored, with the number of failures at the station level ranging from 0 to 34.

When raters were asked about the “attempted” category, many identified using it when something unusual happened during the performance encounter, when a candidate verbalized a task, but did not actually perform it, or when a candidate attempted to complete a task, but it was not done completely or correctly. Raters also noted that when a station scenario was complex and if they found the assessment instruments unclear, the attempted category helped resolve their
discomfort of having to choose between “done” or “not done” as they could choose “attempted” and write a side note to explain. Overall, raters felt that the additional checklist option allowed them to provide a more accurate assessment of each candidate’s current display of medical competence.

An unanticipated, but important, finding that emerged during the qualitative phase of this second study was that of rater differences that may have occurred when raters seemingly disagreed about the fundamental purpose of OSCE assessments. Several raters stated they felt OSCEs should be formative and that it was the responsibility of assessors to provide feedback that learners could use to improve. Several raters also felt the need to give partial credit to trainees. Additionally, some raters felt uninvolved throughout the assessment process (Schumtz, Eppich, Hoffmann, Heimberg, & Manser, 2014), which may have contributed to disagreement regarding the fundamental purpose of the assessment in this instance.

The third study (Chapter 5) - *Mixed Messages or Miscommunication? Investigating the Relationship between Assessors Workplace-Based Assessment Scores and Written Comments* - explored rater variability when assessment occurs in the context of the workplace. Focusing specifically on an Emergency Medicine context, attending physicians used an assessment tool that contained checklist items, rating scales, and written feedback (Chan & Sherbino, 2015) to assess medical residents during their scheduled shifts. What makes workplace-based assessments so challenging is that all residents are assessed for the same core competencies, but each shift represents a unique combination of various cases. Thus, raters’ judgments could be potentially impacted by the fact that each resident is dealing with different patient concerns which allows them to see different things.
One of the main findings from this study was that even when raters were forced to complete an assessment that required them to provide both quantitative and qualitative information, roughly 30% of the assessments were missing narrative comments. This suggests that when observing different things in the workplace, raters do not always feel the need to supplement their quantitative scores with written comments and they will go out of their way to avoid providing such information. One explanation for this could be that raters find it easier and more useful to provide qualitative feedback orally to the resident in the moment (Kogan et al., 2012). Other interpretations may include a lack of time in the workplace context or raters’ unwillingness to commit to something in writing that they may later have to defend.

Upon examination of the relationship between checklist scores, rating scales, and written feedback, it was found that the task rating, but not the global rating, was a significant predictor in determining whether the “done, but needs attention” checklist category was used. This finding not only highlights a relationship between checklists and rating scales, but also suggests that different rating scales are serving different purposes within an assessment instrument.

Another somewhat related finding was that when it comes to determining if a written comment is balanced (i.e., specifies explicit strengths and weaknesses), the use of the “done, but needs attention” checklist category was a significant predictor. When the “done, but needs attention” category was used, raters tended to comment on both positive and negative aspects of a resident’s emerging competence. This suggests that the checklist may be encouraging faculty raters to defend their evaluations by expanding upon what went well and what could be improved in the future.

At the core of postgraduate assessment is competency-based education, in which the goal is to learn and demonstrate competence in specific areas. The respective Royal College of
Physicians and Surgeons in several countries are driving this approach to postgraduate assessment in order to ensure that medical professionals are able to meet the needs of patients in a professional manner (Hodges & Lingard, 2012). Recently, ten Cate (2015) described a false dichotomy of quality and quantity in competency-based assessment in which the quality of learners’ competencies is separated and detached from the quantity of their practical experience. He suggests that this focus on competencies has led some individuals to minimize the value of quantitative features, such as the amount of time spent doing a task or the number of procedures completed, which can be problematic, especially in situations where minimal standards are not met (ten Cate, 2015). This shift may have occurred because for many years quantitative measures were regarded as adequate to demonstrate competence (ten Cate, 2015). Prior to the inception of competency-based medical education, individuals demonstrated competence by successfully carrying out a procedure or task a specified number of times. This current discourse around competency-based assessment is important to consider in the context of rater variance because raters often define competence according to their own experiences, which presently includes those who have experience in quantitative focused or competency-focused assessment systems (ten Cate, 2015). If raters own experiences vary, then their beliefs about how medical competence should best be evaluated may also vary. Therefore, there exists the possibility that the falsely created dichotomy between quality and quantity in competency-based assessment may be also be contributing to the variance observed by raters in performance assessments.

Lastly, findings from this third study support the notion that there exists a “hidden code” (see Ginsburg, Regehr, Lingard, & Eva, 2015) that raters use to communicate their evaluation of residents. It is possible that raters’ differing levels of familiarity with this hidden code is contributing to differences observed in raters’ judgments about an individual’s competence. The
hidden code that raters use may be contributing to errors in communication, rather than simply errors in raters’ judgments. Nevertheless, the presence of a hidden code has broader implications for decision-making, as it would require someone fluent in the hidden code to decipher the overtones and undertones of what raters are attempting to communicate.

Viewing these dissertation findings holistically, and considering them in relation to the theoretical framework outlined in Chapter 2, has important implications for understanding human judgment. Based on the outcomes of the three studies, it would appear as though the dimensionally based categorization perspective is incomplete in explaining how raters formulate impressions and make interpretive judgments. Raters appear to see things idiosyncratically and not always along two orthogonal dimensions. Furthermore, within each of the proposed dimensions, some raters are able to view attributes and abilities more precisely than the perspective suggests. These findings have the potential to influence not only our conceptualizations of how judgments are formulated, but also how human judgment is used to assess competence within medical education.

**Discussion**

From a medical education perspective, competence is defined as a medical professional who is able to practice medicine at a set standard (McGaghie, Miller, Sajid, & Telder, 1978). To further support raters’ understanding of competency-based standards and situate competence within specific medical contexts, entrustable professional activities (EPAs) were developed and employed to assist in making judgments about learners’ emerging competence (ten Cate, 2005). When we rely on raters to make judgments about an individual’s competence (i.e., proficient at a particular level), those judgments are compared to a set standard and the simple act of comparing judgments to standards is what allows the judgments to be deemed better or worse (Baron,
While having standards and allowing raters to make judgments are important aspects of competency-based assessment, the comparison in and of itself introduces variability because raters’ judgments are seen as deviations from a normative standard, which we then refer to as bias (Gingerich et al., 2011).

This research used various quantitative and qualitative approaches to explore how raters’ communicate and articulate themselves, both consciously and subconsciously, when they make evaluative judgments and engage in decision-making. The study, described in Chapter 3, which examined the underlying associations raters make as they formulate evaluative judgments, builds upon the rater-based assessment literature by advancing our understanding of rater idiosyncrasy and forcing us to reconsider the nature of variance in rater-based assessments. Removing the negative stigma associated with rater idiosyncrasy has allowed for more in-depth research that can begin to examine the nuances that exist among individual raters’ judgments (Gingerich et al., 2011). The approach of using multiple analytic methods (e.g., MRFM and hierarchical clustering) in new and innovative ways has allowed for the opportunity to examine complex constructs and hopefully better identify circumstances in which raters’ “biases” are irrational and when they are warranted. This distinction in raters’ differences is crucial in order to determine the validity of raters’ interpretations and judgments in performance assessments.

The discourse of quality and quantity in competency-based assessment has divided quantitative and qualitative measures, each with its own perceived purpose. Many individuals feel that quantitative assessments are about measuring performance in a summative capacity. Despite some of our best efforts to develop more diverse assessment programs (Schuwirth & van der Vleuten, 2012), we still seem to default to quantitative approaches for final decision-making. The strength of such measures lies within their familiarity, which gives the impression that our
current assessment system is functioning well. However, portraying the image that quantitative approaches are adequately meeting our assessment needs leaves little opportunity for change or maintenance in areas where assessment systems could perform better. Some individuals believe that narrative descriptions should replace numerical ratings when assessing clinical performance in medical education (Hanson, Rosenberg, & Lane, 2013). The raters interviewed in Chapter 4 also believed that qualitative assessments would better serve to provide formative feedback that individuals can use to achieve competence in a particular area. Qualitative measures tend to be helpful in addressing the complex aspects of quality in assessment, which may serve to explain why they have been used frequently to justify decisions. An example of complex aspects in which qualitative measures could be useful in aiding to justify a decision can be found in situations where an individual possesses a high level of medical knowledge, but lacks in some of the complimentary skills that are required to fully meet a competency standard (see Figure 2.1 in Chapter 2 for other examples of complex assessments). One of the problems with the modern discourse of assessment is that raters are generally required to provide both quantitative and qualitative judgments in order to make summative decisions and facilitate one’s performance through feedback, which are two distinct and fundamentally different purposes. Perhaps it might be time to ask if we are doing this to ourselves. Is it realistic to have physicians providing feedback, attending to patients, and making high-stakes summative decisions all at once? The “hidden code” described by Ginsburg and colleagues (2015), and further supported by findings from the third study, is a great example of how raters are negotiating multiple purposes and altering assessment instruments used in performance assessments.

Part of what might be observed with rater variance is a product of having dual or multiple purposes for medical education assessments (Townsend, McIlvenny, Miller, & Dunn, 2001). As
illustrated throughout the three studies, raters interact with assessment tools differently. Raters behave in ways that are congruent with their values and beliefs about assessment Yeates, O’Neill, Mann, & Eva, 2013a). Although assessment tools convey performance parameters that raters are expected to attend to, this dissertation illustrates that raters will adapt or modify an assessment tool (or themselves) to meet their own personal needs and style. Therefore, instead of continuing to try and achieve something that may not be feasible or attainable we could utilize the information we currently have about raters in order to use them more purposefully in rater-mediated assessments.

Attempting to untangle the complexity of rater variance has led to a better understanding of individual raters’ strengths and limitations, which could be leveraged to allow for more purposeful rater selection and matching. Raters are independently good at different things. For example, some raters provide excellent feedback, while others can withstand impressive cognitive loads. The ability to identify and support raters in various assessment contexts is crucial in promoting fairness and enhancing the effectiveness of rater-based assessments.

Limitations

This dissertation has explored aspects of raters’ observations and communication in performance assessments within the context of medical education. The three studies mostly represent proof of concept and offer minimal suggestions for how the research findings can be taken further in order to guide the development of a new way of assessing individuals. For example, the first study - Seeing Things Differently or Seeing Different Things? Exploring Raters’ Associations of Non-Cognitive Attributes - offers an opportunity for researchers to identify differences in rating behaviours that could be used to explain how two raters (observing the same performance) can see things differently. However, this study did not offer any practical
suggestions for how these identified differences could be used to better improve and inform the assessment process. What still needs to be further addressed is whether rater idiosyncrasy is a result of raters making meaningful judgments based on their engrained beliefs and experience or if rater idiosyncrasy simply represents uncomplicated measures of CIV. Additional studies such as the ones described throughout this dissertation should continue to monitor rater behaviour over multiple performance assessments in various different contexts.

This dissertation highlights some challenges associated with conducting research in an institutional context. Research carried out at a single institution tends to reflect the current practices of that institution. In the second study, *Understanding the Shades of Grey between “Done” and “Not Done” in Objective Structured Clinical Examination (OSCE) Assessments*, certain decisions were made based upon the current practices of the institution in which the data were collected. For instance, setting the cut score at 60% has certain implications from a standard setting perspective; however, it describes the practices of that particular institution. Other examples such as assignment of candidates and raters to groups, the level in which analyses were conducted, and logistics of rater training are not always within a researcher’s control. Nevertheless, future studies investigating rater variance should ensure that groups are matched based on qualifications of trainees or raters in order to help determine if there are systematic differences between groups.

Another limitation, somewhat related to conducting research in an institutional context, is that this research did not examine the culture of assessment at the various institutions in which these studies were conducted. Culture plays an important role in the assessment process, which can sometimes make it difficult to ascertain the driving force behind raters’ behaviour. Attempting to better understand rater judgments and be more purposeful with information
obtained from assessments, without fully understanding the role that culture plays in various
assessment systems, can make it difficult to promote change, offer recommendations for the
future, or generalize across contexts.

**Implications for Research and Practice**

Despite our best efforts to understand the inner workings and limitations of human
judgment, there is still more research that needs to be done in this area. This dissertation has
focused on the idiosyncratic behaviours and tendencies of raters as individuals in order to
promote assessments that are defensible, valid, and fair for all involved. Closely examining how
raters communicate and articulate their evaluative judgments has highlighted that in order to
develop assessment programs that generate valid results, we need to stop simplifying
individuals’ judgments and trying to force them into discrete categories or orthogonal
dimensions. While each of these dissertation studies explored a different aspect of how
assessment instruments may influence and contribute to rater variance, these findings clearly
suggest a need to revitalize existing assessment practices. One place to start might be to revisit
theoretical frameworks associated with how raters formulate evaluative judgments and consider
the possibility that not all aspects of medical competence can be fully explained by those existing
frameworks (e.g., not all aspects of competence can be adequately defined by two dimensions).

Moving forward, there exists a need not only to figure out how differences observed
among raters can be better reflected and accounted for in high-stakes assessments, but also the
obligation to consider implementing changes when existing assessment systems are creating
barriers for raters. If it is the case that assessment tools are creating barriers for raters, then there
exists the possibility that the tools are also posing unnecessary challenges for students. Such
challenges may include lack of clarity in terms of the purpose for any given assessment, limited
or incomplete feedback, and added stress and anxiety about what is expected to demonstrate competence. Raters, by virtue of their knowledge and experiences, are diverse, bringing a variety of skills, needs, and interests to any assessment situation. To ensure that medical education assessments uphold high, competency-based standards, it might be wise to consider capitalizing on raters’ strengths, attending to their needs, and using the power of human judgment more purposefully. Thus, collecting and using information about individual raters, redesigning assessment tools, and questioning the feasibility of what raters are being asked to do would all serve as valuable steps in improving the existing assessment systems used in medical education.

**Conclusion**

Overall, this research highlights the ‘human’ aspect of human judgments. The first of these dissertation studies showed that raters are not always consciously aware of their behaviour, nor do they possess special insights about how their personal beliefs and biases can impact their judgments and decisions. The second study demonstrated that sometimes raters find it difficult to make what may seem like a simple judgment and there needs to be some sort of mechanism in place to identify when this happens. There is also a responsibility to support raters in making accurate judgments by providing indicators of competence and clearly conveying what is expected at each level of medical training. Further, this study also illustrates that how the information obtained from raters is analyzed and used in decision-making is just as important (if not more important) that the evaluative judgments themselves. Finally, the third study verified that without a certain level of trust and safety in the assessment process, raters will find ways to avoid communicating their true judgments and evaluations. Medical education assessment systems need to put trust in their raters and the clinical judgments that these qualified medical professionals can provide. This includes situations beyond a specific assessment encounter where
it might be necessary to uphold difficult decisions such as failing someone who does not meet the standards. The findings presented in this dissertation stress the importance of establishing assessment programs that can support raters’ subjective judgments and capitalize on the strengths of individuals with specific expertise. Only then will it be conceivable to take advantage of human judgment and uphold the principles of fairness, transparency, and defensibility in performance-based assessments.
REFERENCES


doi:10.1111/j.1365-2929.2006.02692.x


H Hamilton Health Sciences  
McMaster University  
St. Joseph’s Healthcare Hamilton

Hamilton Integrated Research Ethics Board

Suzette Salama, PhD.
293 Wellington Street North
Suite 102
Hamilton, On
L8L 8E7

May 28, 2014

To Whom It May Concern:

Please accept this letter to clarify that the McMaster Modular Assessment Program (McMAP) quality assurance program has been granted an exemption from the Hamilton Integrated Research Ethics Board (HIReB). The HIReB is a joint, city-wide initiative that administers the institutional review and research ethics proceedings for McMaster University, Hamilton Health Sciences Corporation, and St. Joseph’s Healthcare.

According to the Tri-Council Policy Statement 2, Chapter 2, we have previously acknowledged the review and scholarly work around the McMAP is deemed to be Quality Assurance and Improvement, since the results of any work done in this area has previously been directly used to evaluate and/or improve an existing and implemented educational program within the McMaster University Emergency Medicine Residency programs.

Dr. Teresa Chan has been leading the McMAP quality assurance project, and has now notified me that she will be inviting an external expert (Stefanie Sebok, PhD Candidate) to continue in the active and ongoing evaluation of the program. Ms. Sebok will be assisting Dr. Chan in these activities, and will be specifically focusing on validity testing for the McMAP clinical observational checklists, scores and qualitative commentaries. According to Dr. Chan, Ms. Sebok’s work will be a much-welcomed addition to the ongoing project improvement.

Please do not hesitate to contact me if you have any further issues or concerns.

Yours truly,

S. Salama
Suzette Salama
Co-Chairperson,
Hamilton Integrated Research Ethics Board