A MANY-FACET RASCH MEASUREMENT ANALYSIS TO EXPLORE RATER EFFECTS AND RATER TRAINING IN MEDICAL SCHOOL ADMISSIONS

by

Amanda Brijmohan

A thesis submitted in conformity with the requirements for the degree of Master of Arts
Department of Leadership, Higher and Adult Education
Ontario Institute for Studies in Education
University of Toronto

© Copyright by Amanda Brijmohan 2016
A MANY-FACET RASCH MEASUREMENT ANALYSIS TO EXPLORE RATER EFFECTS AND RATER TRAINING IN MEDICAL SCHOOL ADMISSIONS

Master of Arts 2016
Amanda Brijmohan
Department of Leadership, Higher and Adult Education
University of Toronto

Abstract

For many undergraduate medical education (UME) programs, ratings of applicants’ performance in interviews are integral to the decision making process. This study used many-facet Rasch measurement (MFRM) to explore the severity and consistency of 200 interviewers’ ratings of 599 applicants to a UME program. The raters varied widely in their severity and many of the raters had poor fit to the MFRM model. Although raters at four stations rated both station-specific and across-station competencies, the highest correlations were within station; the correlations across stations for the common competencies were low. No significant relationships were found between raters’ training mode(s) (online and in-person) and rating behaviours. The findings from this work demonstrate the need to make explicit how differences in rater judgement affect fairness of applicant assessment to UME programs.
ACKNOWLEDGMENTS

Behind the production of this thesis was a network of support that any researcher, student, friend, sister, or daughter would be blessed to have, and it is to them that I owe my deepest gratitude. Thank you to Dr. Ruth Childs for your guidance, your patience, and your kindness during the course of supervising this thesis. Thank you for introducing me into the world of measurement, a field that consistently challenges me, both as a student, and as a researcher. Thank you to Dr. Leesa Wheelahan for your unwavering support and encouragement; I’ve grown immensely over the past two years under your research team, both academically and professionally; it is to your efforts that I owe thanks.

This work would not have been made possible without the support of Dr. Mark Hanson, & Hana Lee at the Faculty of Medicine/U of T. Thank you for providing an opportunity from which to explore the mysteries that underpin rater cognition.

Thank you to my OISE colleagues and friends for your support, and in sharing my excitement as I pursued this academic journey.

Most importantly, thank you to my family. To my father, who inspires me every day to work hard, and earn what I have. To my mother, whose unconditional love made even the most difficult days of writing that much easier. To my sister, who consistently proves to be the perfect combination of both. None of this would have been possible without you; accomplishing this milestone is as much your success as it is mine. For that, I thank you.
# TABLE OF CONTENTS

Chapter 1. Introduction ........................................................................................................... 1

Chapter 2. Literature Review ................................................................................................. 3
  Research Context .................................................................................................................. 3
  Part I: Admissions in Undergraduate Medical Education .................................................. 4
    Canadian UME Admissions Tools used in Acceptance Decisions .................................. 4
    Variations of the Undergraduate Medical School Interview .......................................... 5
  Part II: Rater Effects .......................................................................................................... 8
    The Rating Process ........................................................................................................... 9
    Rater Training .................................................................................................................. 11
    Evidence of Rater Effects in Assessment in Medical Education .................................... 12
  Part III: Measurement Theory .......................................................................................... 13
    Classical Test Theory ...................................................................................................... 13
    Item Response Theory .................................................................................................... 15
  This Study .......................................................................................................................... 15

Chapter 3. Methods ................................................................................................................. 16
  Data .................................................................................................................................... 16
    The Modified Personal Interview .................................................................................. 16
    Competencies .................................................................................................................... 18
    Rater Training .................................................................................................................. 20
  Data Analysis ...................................................................................................................... 20

Chapter 4. Results .................................................................................................................... 21
  Research Question 1: What are the distributions of ratings and the correlations among the
    ratings of competencies across interview stations? .......................................................... 21
  Research Question 2: How well is the rating behaviour modelled by the MFRM? How lenient
    and severe are the raters? .............................................................................................. 25
  Research Question 3: How are modes of training related to rater behaviour? .................... 32

Chapter 5. Discussion and Conclusion ................................................................................... 36
  Limitations of the Study and Future Directions .................................................................. 39

References ................................................................................................................................ 41
LIST OF TABLES

Table 1 Applicant Interview Cycle ........................................................................................................ 17
Table 2 Rater Interview Cycle .................................................................................................................. 17
Table 3 Distributions of Ratings by Station .............................................................................................. 22
Table 4 Distributions of Raters’ Mean Ratings Across Stations ................................................................. 23
Table 5 Correlations of Ratings Across Competencies .............................................................................. 24
Table 6 Raters’ Participation in Training ..................................................................................................... 32
Table 7 Raters’ Participation in Training by Station ...................................................................................... 33
Table 8 Average Theta for Raters who Participated in Different Modes of Training by Station. 34
Table 9 Average Absolute Infit for Raters who Participated in Different Modes of Training by Station .................................................................................................................. 35
LIST OF FIGURES

Figure 1. Applicants’ mean ratings against theta-A. ............................................................... 26
Figure 2. Applicants’ infit mean squares against applicants’ ability estimates (theta-A). ........ 27
Figure 3. Raters’ mean ratings against theta-R. ................................................................... 28
Figure 4. Applicants’ infit mean squares against applicant ability (theta-R). ...................... 29
Figure 5. Wright Map of raters, applicants, stations, and MPI scores. ............................... 31
CHAPTER 1. INTRODUCTION

Admission decisions by Canadian Undergraduate Medical Education (UME) are high-stakes for applicants, institutions and society. With notably low attrition rates in UME, Residency, and throughout the profession (Eva & Reiter, 2004), admission decisions will largely determine who will make up our future healthcare workforce. Future physicians will ultimately effect the health and well-being of the populations they serve; therefore, it is the social and ethical responsibility of Medical Education institutions to not only train competent physicians, but to recognize and select applicants who stand the best chance of developing into competent physicians (Patterson & Ferguson, 2010; Patterson et al., 2016). Secondly, Canada has been recognized as a global leader in medical education research, teaching, and practice, thus attracting a large and highly qualified applicant pool of domestic and international top-talent (The Association of Faculties of Medicine of Canada, 2012). With such competition for limited seats, effective and reliable admission processes are essential. For the interview stage in medical admissions, especially, it can be difficult to test claims of validity and reliability. In pursuit of more reliable measures, relatively new psychometric analytic techniques may help to shed light on what unconsidered factors affect the ways in which applicants are rated; of particular interest are ways in which raters affect the measurement process. Using existing data from a UME admission process, this study will examine rater effects by employing many-facet Rasch measurement (MFRM) to explore rater severity, leniency and consistency. As well, this study will relate different kinds of rater training (online (prior), in-person (prior), day of) with systematic differences in scoring behavior as shown from the MFRM analysis.

Chapter 2 provides a description of the research context, and discusses the literatures on rating and rater effects, relevant measurement theory, and the development of the medical school
Chapter 3 describes the datasets used in this study: the applicant interview scores from a large UME program, and the rater training dataset. The results are presented in Chapter 4, in three parts: part 1 will discuss frequency distributions and correlations among rater and station means to examine any emerging trends in rater behavior. Part 2 will present the results from the multi-facet rasch model and will examine how the estimated parameters relate to actual ratings given to applicants. Part 3 will present chi-square statistics and analysis of variances to examine if rater training modes are related to rating behaviours. Chapter 5, the Discussion and Conclusion, will provide an interpretation of the findings, and discuss implications, future directions, and limitations of this study.
CHAPTER 2. LITERATURE REVIEW

Research Context

The 21st century has seen the Canadian Healthcare system evolve within a dynamic societal, political, and economic climate. With increasing globalization, comes a diverse populace, with even more diverse healthcare needs. To be able to navigate such a landscape, not only must the modern-day physician respond in adapting their practice and delivery of care, but so too must the field of medical education respond in training and selecting students who are capable of meeting present-day challenges. This idea is not novel; the notion of the system of medical education being responsive to societal needs was inherent within the reforms first suggested by Abraham Flexner over a century ago. First commissioned by the Carnegie Foundation for the Advancement of Teaching, Abraham Flexner performed an evaluation of 155 medical schools across North America, the findings of which were published in his report “Medical Education in the United States and Canada (1910).” In his review, Flexner drew attention to the need for standardization of admissions requirements, the need for curricular reform, and the important role that academic medicine had to play in the quality of care.

Throughout the remainder of the 20th century, Flexnerian reforms have had a lasting effect on medical education in North America. Most notable was the recognition that medical education had a responsibility to society in understanding that the faculties of medicine’s selection decisions would ultimately determine the composition of the future healthcare workforce (World Health Organization, 1996). Attention to this relationship has also been recognized as a priority of research within the Canadian context. At the turn of the new millennium, the Association of Faculties of Medicine of Canada (AFMC) took part in
establishing the “Future of Medical Education in Canada (FMEC)” project which made suggestions for improving the current Canadian Medical Education system. Highlighted in their recommendations was a re-examination of systemic issues in the selection practices of medical school applicants and a call to focus research efforts on making admissions practices fair, transparent, and equitable to all populations who apply (Bandiera, Maniate, Hanson, Woods, & Hodges, 2015). As well, the 2010 Ottawa Conference on the assessment of selection and training of health professionals issued a consensus statement which called for more research on the psychometric refinement of admissions tools (Prideaux et al., 2011). In response to these calls, there has been a growing body of research devoted to examining admissions to medical school; questions are emerging regarding what traits we value in our health care practitioners, if we fairly recognize these traits in our applicants, and whether or not our admissions processes are equitable, and our decisions on who gets accepted defensible.

This chapter reviews three literatures that are important for this study: admissions especially interview development in medical education; rater effects; and measurement theory.

**Part I: Admissions in Undergraduate Medical Education**

**Canadian UME Admissions Tools used in Acceptance Decisions**

UME admissions use a number of tools to evaluate applicants on their cognitive and non-cognitive traits. Cognitive traits refer to a applicant’s academic ability, represented by their previous undergraduate Grade Point Average (GPA) in combination with scores from their Medical College Admission Test (MCAT). Ubiquitously relied upon by Canadian UME institutions, applicants’ GPAs and MCAT scores have long been relied upon in the admissions decision, as these metrics have demonstrated the strongest predictive validity of medical student
academic performance and licensing examinations (Donnon, Paolucci, & Violato, 2007; Kulatungo-Moruzi & Norman, 2002; Salvatori, 2001). However, though academic ability is necessary for medical school success, it cannot be assumed that academically high-performing students will become good doctors (Patterson & Ferguson, 2010). As well, there is growing recognition of the importance of non-cognitive traits in physician competence; research suggests that non-cognitive traits not only predict success in medical school, but also success in clinical practice, and improved patient outcomes (Grumbach & Bodenheimer, 2004; Koenig et al., 2013). With this recognition, there is a shift in the ways non-cognitive traits are viewed, measured and weighted during admissions processes (Eva & Reiter, 2004; Koenig et al., 2013; Patterson & Ferguson, 2010; Patterson et al., 2016). Non-cognitive traits (also called personal competencies, interpersonal characteristics, personal qualities, and non-cognitive dimensions) refer to the applicant’s interpersonal attributes, social skills, interests, and motivations. Examples of desired non-cognitive traits would be strong communication, professionalism, empathy, adaptability, and ethical decision-making (Eva, Rosenfeld, Reiter, & Norman, 2004; Koenig et al., 2013). Within the admissions process, non-cognitive traits are assessed through several modalities; to inform their decisions, admissions committees rely on their judgments of the applicant’s autobiographical statement, reference letters, and most of all, their medical school interview.

**Variations of the Undergraduate Medical School Interview**

The medical school interview as an assessment tool has been relied upon as the richest form of non-cognitive evaluation for applicants in the UME admissions process. Edwards, Johnson, and Molidor (1999) note that the interview serves four important purposes: information gathering, verification of applicant data, informing decision making, and as a recruitment
opportunity for the medical school (Edwards, Johnson, & Molidor, 1990). The original way interviews were administered (often referred to as traditional, panel, or personal interview in the literature) consisted of a single station, lasting roughly 45 minutes in length, in which the applicant responded to items posed by a panel of interviewers. The personal interview would either be structured or semi-structured, and items would range from questions about past behaviour to situational judgment tasks (Edwards et al., 1990; Eva, Rosenfeld, et al., 2004). Though institutions rely heavily on the personal interview to inform their acceptance decisions, the psychometric integrity of the format has consistently been called into question (Pau et al., 2013; Siu & Reiter, 2009).

**Issues with the Personal Interview Format**

Research has repeatedly shown that, as an assessment tool, the personal interview demonstrates high variability in reliability estimates, and poor predictive validity (Kreiter & Axelson, 2013; Pau et al., 2013). Eva and colleagues (Eva, Neville, & Norman, 1998; Eva, 2003) argue that context specificity in the personal interview structure is a contributor to its poor reliability. Context specificity suggests that skills and behaviors have less to do with the latent traits of the individual, and more to do with the context of the situation (Eva et al., 1998; Eva, 2003). In the personal interview, context specificity becomes problematic, as generalizability of the interview scores becomes compromised. Findings by Kreiter et al. (2004) show that within the personal interview, the variance attributed to the “participant × occasion” interaction was greater than that attributed to the differences between the participants themselves (Kreiter, Yin, Solow, & Brennan, 2004). These findings are problematic, as the personal interview becomes more a measure of the situation the applicant happens to be in, rather than the applicant.
Multiple Independent Sampling and the Multiple Mini-Interview

One way to limit the effects of context specificity on reliability is if the behavior is assessed on multiple occasions. For example, if the applicant demonstrates strong communication skills on multiple occasions, there is confidence in saying that the applicant is an effective communicator. In many domains of performance assessment, this principle is what underpins the Multiple Independent Sampling (MIS) method. MIS allows for the assessment of performance on multiple occasions, using different raters conducting independent evaluations (Hanson, Kulasegaram, Woods, Fechtig, & Anderson, 2012).

Within the context of medical school admissions, the MIS method has been incorporated into the Multiple-Mini Interview (MMI). Developed out of McMaster University, the MMI is a structured interview format in which applicants rotate through a series of 10 interview stations designed to assess their non-cognitive traits. Applicants are to read a scenario or question and then discuss their response with interviewers who may be a faculty member, a medical student, or not medically affiliated. Applicants are then assessed on their critical thinking, ethical decision making, communication skills, and knowledge of the healthcare system (Eva, Rosenfeld, et al., 2004; Pau et al., 2013). The MMI has demonstrated improved reliability (Eva, Rosenfeld, et al., 2004; Pau et al., 2013) and has shown promising predictive validity (Eva, Reiter, Rosenfeld, & Norman, 2004; Eva et al., 2009). However, regardless of these psychometric improvements, there has been resistance to the wide-spread adoption of the MMI method. Hanson et al. (2012) notes three possible reasons for this: the MMI imposes high costs, reduces recruitment opportunity for the institution, and lacks flexibility in the design (Axelson & Kreiter, 2009; Hanson et al., 2012).
The Modified Personal Interview

In trying to maintain the benefits of both interview formats, Axelson and Kreiter’s (2009) research showed that the MIS method could be applied to the personal interview. Their findings suggest that the MIS method can improve reliability by decreasing the number of interviewers, while increasing the number of stations (Axelson & Kreiter, 2009). Building on these findings, Hanson and colleagues (Hanson et al., 2012) developed a new interview protocol called the Modified Personal Interview (MPI). Developed out of the Faculty of Medicine at the University of Toronto, the MPI requires applicants to rotate through four consecutive personal interviews, each lasting 10-12 minutes, in which they are asked questions about past behaviours. Generalizability and decision studies subsequently showed marked improvements in reliability. Hanson et al. (2012) note that these findings suggest a way for admissions committees that favour personal interviews to maintain reliability while also balancing resourcing requirements and recruitment efforts by institutions (Hanson et al., 2012).

Though these findings suggest the MPI may improve the measurement of the personal interview, the structure of the MPI only addresses the variance in reliability due to context specificity effects. Less understood are the effects that interviewers/raters have on the interview process; as the interview is a rater-mediated assessment, it is important to study how interviewers as raters contribute to variance in scores assigned to applicants, as well as how rater effects influence the overall reliability of the measurement.

Part II: Rater Effects

When assessing applicants through interview methods, there are many sources of variance that affect what score the applicant is given. For informing their acceptance decisions, what is most useful to interviewers/raters is the variance that represents meaningful differences
between the applicants (high ability applicants vs. low ability applicants). What is least useful is the variance in scores due to extraneous or irrelevant factors. This kind of variance makes it difficult to distinguish between applicants, and consequently, threatens the fairness of the measurement, and meaningful interpretation of results (Eckes, 2011; Lane & Stone, 2006; McNamara & Roever, 2006; Weir, 2005). Interestingly, raters themselves may contribute to the latter source of variance. The following will review how raters have been conceptualized as a subject of study within the field of measurement. It is important to note that this literature demonstrates inconsistency in the terminology which describes rater variability. It has been called construct-irrelevant variance (Haladyna & Downing, 2004; Messick, 1989), rater error (Downing, 2005; Saal, Downey & Lahey, 1980), rater bias (Hoyt, 2000; Johnson, Penny, & Gordon, 2009), and more collectively, rater effects (Wolfe & Myford, 2003; Scullen, Mount, & Goff, 2000).

This paper will use the definition from Scullen, Mount, and Goff (2000), who defined rater effects as a “broad category of effects (resulting in) systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee” (p. 957).

**The Rating Process**

When rating a task, skill, performance, or applicant, Myford and Wolfe (2003) note that raters are called upon to act as information processors, in that they (either concurrently or consecutively) “engage in a highly sophisticated, complex mental process to arrive at their decisions—observing, recalling information from memory storage, and then organizing, combining, weighing, and integrating that information to draw inferences about individuals” (Myford & Wolfe, 2003, p. 387). Thorndike and Hagen (1977) also describe the rating process as
“an evaluative summary of past or present experiences in which the ‘internal computer’ of the rater processes the input data in complex and unspecified ways to arrive at the final judgement” (Thordike & Hagen, 1977, p. 449). Eckes (2011) describes the process of rating as a “long, and possibly fragile, interpretation-evaluation-scoring chain” of events in which the rater makes a judgement about the trait being measured, all the while building on his/her own understanding of the construct and adhering to scoring rubrics and guidelines.

Though research in this area has often characterized the rater as a “computer” or an “information processor,” scholars have long acknowledged that the rating process is far from systematic. Due to idiosyncratic differences in rater cognition, each rater will see, interpret, and evaluate the same performance differently (Eckes, 2011; Gingerich, Kogan, Yeates, Govaerts, & Holmboe, 2014; Gingerich, Regehr, & Eva, 2011). Raters have been shown to differ not only in their evaluation of performance, but also in their own understanding of the construct being measured, how they interpret scoring criteria, and how they interpret rater training (Eckes, 2011; Lumley & McNamara, 1993; Myford, 2003). These differences manifest as rater variance, which interferes with both the reliability and the validity of the interview as a measurement tool (Eckes, 2011; Sebok & Syer, 2015).

Some examples of rater effects that have been well studied (see Myford & Wolfe, 2003, for a comprehensive review) are rater leniency, severity, halo effects and central tendency effects. Rater leniency refers to the tendency of a rater to systematically award high ratings; this is typically seen as a rater awarding higher ratings than other raters and tending to use the top part of the rating scale. Rater severity, in contrast, refers to the tendency of a rater to award lower ratings. Central tendency is when raters avoid the high and low ends of the rating scale. The halo
effect occurs when a rater focuses on one attribute of an applicant and then rates other attributes similarly.

**Rater Training**

Rater training has been commonly used in trying to reduce rater variance; the assumption is that training raters on how to perceive performance, as well as interpret and use criteria, should improve inter-rater reliability. However, previous research suggests that rater training seems to have little effect (Cook, Dupras, Beckman, Thomas, & Pankratz, 2009; Eckes, 2011; Gingerich et al., 2014, 2011; Holmboe, Hawkins, & Huot, 2004; Sebok & Syer, 2015). Within the context of medical education, rater training interventions have been investigated in the assessment of residency performance using the mini-clinical evaluation exercise (mini-CEX). However, rater training did not seem to have an effect on increasing reliability or accuracy of scores (Cook et al., 2009). One explanation for these results is that some raters are seen as trainable (Gingerich et al., 2014) while others are “impervious to training” (Williams, Klamen, & McGaghie, 2003). Another reason is that even for the most seasoned raters who have gone through extensive training, heterogeneity of their scoring patterns still arises, as raters may focus on different aspects of the performance which influence how they score (Eckes, 2011; Murphy, Cleveland, Skattebo, & Kinney, 2004). Designing the best possible rater training for the MPI requires a greater understanding of how these rater effects respond to different kinds of training modes (e.g., in-person, online), timing (e.g. immediately before the interviews, and/or days or weeks earlier), and duration. Additional research investigating the relationship between these design decisions and rater variance is needed.
Evidence of Rater Effects in Assessment in Medical Education

Rater effects have consistently been acknowledged as a problem within rater-mediated assessments used in medical education. Research has consistently shown that raters’ scoring patterns during the evaluation of the same performance can vary quite dramatically. Holmboe et al. (2003) investigated the construct validity of the miniCEX. In this study, faculty members were to watch and rate the performance of actors playing standardized residents using a 9-point rating scale. Findings showed that raters demonstrated poor inter-rater reliability, with ratings varying as much as 6 points.

The objective structured clinical examination (OSCE) is another assessment approach in which rater variability has been well studied. In the examination of how raters assess professionalism in the OSCE, Mazor et al. (2007) showed that raters varied dramatically in how they attended to both specific attributes of performance and in their global impressions. In 19 of the 20 OSCE stations, the study found between one and eight discrepancies in judgement, in which one rater thought an observed behaviour was positive, whereas another perceived it as negative. In the context of medical school admissions, rater variability has also been investigated as an issue to consider during the decision making process. In examining the reliability and validity of the MMI for selection of applicants for graduate-entry medical school, Roberts et al. (2008) found that variation in rating was largely due to rater subjectivity; as well, the correlations of scores of applicants across stations were low. These findings collectively suggest the need to further explore the cognitive underpinnings which guide how raters arrive at their judgements, and to investigate what factors directly and indirectly affect their scoring patterns.

Recently, these issues have been investigated using relatively new psychometric techniques. Sebok and Syer (2015) employed a combination of statistical models
(generalizability theory, MFRM and hierarchical clustering) in pursuit of understanding how rater variance affects the MMI. Findings from their MFRM analysis are of particular interest, as they demonstrate how such analyses can provide valuable insight into rater effects on interview scores. MFRM results from the Sebok and Syer (2015) study identified problematic raters, while also showcasing how statistical methods could allow a deeper investigation of the idiosyncrasies of rater judgment which can manifest as rater variability in scores. These kinds of psychometric techniques are powerful in helping to understand how the MPI could be optimized for measurement of applicant ability. To provide some background, the next section will briefly touch on dominant measurement theories, as well as introduce many-facet Rasch measurement (MFRM), the main technique used in our analysis.

**Part III: Measurement Theory**

In measurement, a test score, or in this case, an admissions interview score is assumed to reflect the ability of the applicant, or his/her performance in the trait being assessed. However, as seen from the rater effects literature, we can no longer be confident in that assumption; the scores assigned to interview applicants may not entirely reflect their ability, but may also reflect the rating behaviours of their assigned raters. This highlights the need to better understand two things: (1) the relationship between the scores assigned to applicants and their actual ability and (2) how rater effects distort this measurement process. Test theory models provide a way of investigating that relationship. The following literature review will briefly cover dominant test theory models and how we arrived at MFRM as most appropriate for investigating our data.

**Classical Test Theory**

First put forth by Charles Spearman (1904: General Intelligence, Objectively Determined and Measured) and later explained by Lord and Novick (1968), Classical Test Theory (CTT)
assumes that when an individual is given a test score, that test score represents a combination of the individual’s true score and measurement error (Rust & Golombok, 2009). This can be seen by the following equation:

\[ X = T + E \]

In this equation, \( X \) is the observed score, \( T \) is the true score, and \( E \) is the measurement error. A person’s true score is the most accurate proxy of their true ability on what is being tested; this true score is what the person would have gotten after an infinite number of test administrations, in which case the measurement error would be zero (De Champlain, 2010). In such a situation, the observed score is assumed to equal the true score. Practically speaking, an infinite number of test administrations, (or in this case, an infinite number of interviews) would not be feasible. Therefore, CTT allows for a reliability coefficient to be calculated which provides an estimate of how close the observed scores are to the applicant’s true score.

**Limitations of Classical Test Theory**

For the purposes of psychometrically analyzing the interview process, CTT as a measurement framework has some notable limitations. First, CTT only takes into account how the applicants perform in relation to the difficulty of the question and vice versa. This model would not separate the variance due to differences in applicant ability outside of the difficulty of the questions (Tavakol & Dennick, 2013). Even more important, this model would not allow for an examination of the variance due to raters; rather this variance would just be aggregated within the error term. Secondly, the model assumes that measurement error is the same across all ratings, which may not be true (Hambleton & Jones 1993).
Item Response Theory

Item response theory (IRT), in contrast to CTT, estimates the probability of a response or rating in relation to an underlying continuum of ability and does not assume that reliability is constant across the continuum. An advantage of IRT is that applicant ability and item difficulty are estimated independently from each other (Andrich, 2004). MFRM is an IRT model that, in addition to estimating examinees' or applicants' ability and items' or traits' difficulty, also estimates raters' severity (Linacre, 1989). MFRM has the advantage of placing all these estimates on the same ability continuum.

This Study

To date, there has not been any study using MFRM to study rater effects on the MPI format. As well, how rater effects are mediated by different kinds of rater training has not been examined. This current study seeks to fill both gaps.

Using existing data from the admission process for the Undergraduate Medical Education program at the University of Toronto’s Faculty of Medicine, this study uses MFRM to investigate rater effects and relate those effects to rater training. The research questions are:

1. What are the distributions of ratings and the correlations among the ratings of competencies across interview stations?

2. How well is the rating behaviour modelled by the MFRM? How lenient and severe are the raters?

3. How are modes of training related to rater behaviour?
CHAPTER 3. METHODS

Data

This study employed a retrospective quantitative analysis of interview data from the admissions department of the University of Toronto Faculty of Medicine’s Undergraduate Medical Education (UME). The data comprised the interview scores of the applicant pool applying for admission to the Faculty of Med/U of T for Fall 2016. The dataset included applicant interview scores, as well as anonymized applicant and rater IDs. From February to April of 2016, 599 applicants were interviewed by 200 raters. Each rater was assigned to 1 station, and rated 12 applicants (refer to Table 1). The rater group included physicians, medical students, residents, and other health care professionals. The admissions department also provided two datasets pertinent to rater training. The first dataset summarized raters’ portal usage as a proxy for participation in online-training, and the second summarized raters’ attendance to the in-person training sessions.

The Modified Personal Interview

In 2013/2014, the University of Toronto Undergraduate medical education admissions interview was redesigned into the Modified Personal Interview (MPI). Each MPI circuit consists of four independent semi-structured interviews (each lasting 12 minutes) that applicants rotate through. Each interview is conducted by a single rater; raters are given a set of pre-determined questions for each station to pose to applicants, and are allowed to ask their own questions as well. After the interview, raters are given three minutes to record their ratings and notes for each applicant. The following tables outline the interview cycle from an applicant’s perspective (Table 1) and from a rater’s perspective (Table 2).
### Table 1

**Applicant Interview Cycle**

<table>
<thead>
<tr>
<th>Station 1: Collaboration (1A)</th>
<th>Station 2: Ethical Decision Making (2A)</th>
<th>Station 3: Reflection (3A)</th>
<th>Station 4: Values (4A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maturity (1B)</td>
<td>Maturity (2B)</td>
<td>Maturity (3B)</td>
<td>Maturity (4B)</td>
</tr>
</tbody>
</table>

Communication and Interpersonal Skills (1C)
Communication and Interpersonal Skills (2C)
Communication and Interpersonal Skills (3C)
Communication and Interpersonal Skills (4C)

Caring (1D)
Caring (2D)
Caring (3D)
Caring (4D)

Applicant rotates through 4 stations in a randomized fashion. Each station lasts 12 minutes; the entire interview process lasts 60 minutes.

### Table 2

**Rater Interview Cycle**

<table>
<thead>
<tr>
<th>Timeslot</th>
<th>Station 1 (Rater 1)</th>
<th>Station 2 (Rater 2)</th>
<th>Station 3 (Rater 3)</th>
<th>Station 4 (Rater 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Applicant 1</td>
<td>Applicant 2</td>
<td>Applicant 3</td>
<td>Applicant 4</td>
</tr>
<tr>
<td>2</td>
<td>Applicant 4</td>
<td>Applicant 1</td>
<td>Applicant 2</td>
<td>Applicant 3</td>
</tr>
<tr>
<td>3</td>
<td>Applicant 3</td>
<td>Applicant 4</td>
<td>Applicant 1</td>
<td>Applicant 2</td>
</tr>
<tr>
<td>4</td>
<td>Applicant 2</td>
<td>Applicant 3</td>
<td>Applicant 4</td>
<td>Applicant 1</td>
</tr>
<tr>
<td>5</td>
<td>Applicant 5</td>
<td>Applicant 6</td>
<td>Applicant 7</td>
<td>Applicant 8</td>
</tr>
<tr>
<td>6</td>
<td>Applicant 8</td>
<td>Applicant 5</td>
<td>Applicant 6</td>
<td>Applicant 7</td>
</tr>
<tr>
<td>7</td>
<td>Applicant 7</td>
<td>Applicant 8</td>
<td>Applicant 5</td>
<td>Applicant 6</td>
</tr>
<tr>
<td>8</td>
<td>Applicant 6</td>
<td>Applicant 7</td>
<td>Applicant 8</td>
<td>Applicant 5</td>
</tr>
<tr>
<td>9</td>
<td>Applicant 9</td>
<td>Applicant 10</td>
<td>Applicant 11</td>
<td>Applicant 12</td>
</tr>
<tr>
<td>10</td>
<td>Applicant 12</td>
<td>Applicant 9</td>
<td>Applicant 10</td>
<td>Applicant 11</td>
</tr>
<tr>
<td>11</td>
<td>Applicant 11</td>
<td>Applicant 12</td>
<td>Applicant 9</td>
<td>Applicant 10</td>
</tr>
<tr>
<td>12</td>
<td>Applicant 10</td>
<td>Applicant 11</td>
<td>Applicant 12</td>
<td>Applicant 9</td>
</tr>
</tbody>
</table>
Competencies

Each station assesses a station-specific competency, and three competencies common across all four stations. The four station-specific competencies are “collaboration,” “ethical decision making,” “reflection” and “values.” The three common competencies are “maturity,” “interpersonal skills” and “caring.” The following sections will explore each competency in further depth, and describe how raters were to assess applicants on these competencies.

Station-specific Competencies

Collaboration (1A). At this station, raters were to ask questions of applicants that explored their understanding of collaboration as a necessary quality of physicians for providing holistic healthcare. The questions posed to applicants were intended to encourage them to recount their experiences in team-based activities (either professional, academic or extra-curricular activities). Applicants were also asked to speak about their experiences working in inter-professional health teams, with attention to how they reflected on their role.

Ethical Decision Making (2A). At this station, raters were to identify applicants who demonstrated sound judgement and the capacity to recognize and deal with ethical issues in healthcare. Applicants were given two ethical decision making scenarios. From their responses, raters were to gauge whether applicants are able to identify the ethical conflicts, to view the scenario with different frames of reference, and to reflect on the consequences of healthcare decisions made.

Reflection (3A). For the reflection station, raters were specially recruited, and were trained to use the REFLECT rubric (Wald, 2012). Raters were to read applicants’ brief personal essays and construct their own questions regarding reflective capacity. Raters were to examine
whether applicants could analyze, question and reframe experiences in order to improve their own practice or the health care of the patient.

Values (4A). Within the values station, raters were to explore how applicants’ underlying belief systems related to their pursuit of a medical career. Raters asked applicants about their values, knowledge, skills and experiences that reflected the applicants’ valuing of integrity, social justice, equity, and life-long learning. The goal of the rater was to explore whether applicants demonstrated potential to develop into caring healthcare professionals.

Common Competencies (Across All Stations)

Maturity (1B, 2B, 3B, 4B). Raters were to assess the applicants’ overall level of maturity demonstrated in their interview responses. Indicators included the applicants’ emotional, social, and intellectual maturity. Using applicants’ recounted experiences, raters were to gauge whether applicants demonstrated the knowledge, and motivation needed for a career in medicine.

Communication and Interpersonal Skills (1C, 2C, 3C, 4C). Raters were to assess whether applicants could articulate their responses in a clear and professional manner. Raters were to pay attention to both verbal and non-verbal communication skills, as well as whether the applicants were good listeners. Raters were to take note of the applicants’ interpersonal skills and self-confidence in how they communicated their responses.

Caring (D). Raters were to give a global assessment of applicants’ potential to develop into caring physicians. Indicators raters looked for are if applicants demonstrated the capacity for empathy, compassion, and respect needed for healthcare.

Rating Scale

Raters used a 7-point Likert-type scale for their ratings of applicants’ competencies. The points were labeled as follows:
1  2  3  4  5  6  7
Unsuitable   Acceptable   Good   Very Good   Excellent   Outstanding   Superb

Raters were also to award a green flag rating to applicants who should be offered admission, and a red flag rating to those who should not, followed by an explanation of their reasoning.

**Rater Training**

Raters were provided three modes of training: Online training (optional), in-person training (optional), and day-of training (required).

**Online Training**

For each station, raters were provided with an overview video, in which the entire interview process was outlined. Raters were also provided with training modules for each station, and two online simulation videos, used to assist in their line of questioning, as well as outlining their role during the interview process. For the reflection station, additional training videos for content relevant to the assessment of reflective capacity were provided.

**Data Analysis**

Descriptive statistics were computed using Excel 2015 and SPSS 23. For the many-facet Rasch measurement (MFRM) analysis, FACETS for windows Version No. 3.71.4 was used. A three-facet Rasch model was defined (for raters, applicants, and station). The applicant facet was centered with a standard deviation of 1. Chi-square analyses and analyses of variance were run using SPSS to examine whether training mode had any significant difference between stations, or any relationship with ratings or with the infit estimated by FACETS.
CHAPTER 4. RESULTS

The results are presented in three sections. First, frequency distributions and correlations among the ratings are presented for both raters and stations. Second, results from the many-facet Rasch measurement (MFRM) analyses are presented. This will include the Wright Map, as well as scatterplots depicting the infit statistics. Third, chi-square tests and analyses of variance will be used to compare the frequencies and ratings across stations and training modalities. Recall that the data for this analysis consists of the MPI ratings of 599 applicants. The applicants were divided into groups of 12. Each group of 12 went through four stations (applicants could begin at any of the four stations), with four competencies per station. Each station had a different rater and each rater was assigned to only one station. Each applicant was rated on a scale from 1 to 7 for each competency within each station, for a total of 16 ratings per applicant.

Research Question 1: What are the distributions of ratings and the correlations among the ratings of competencies across interview stations?

Table 3 shows the percentage of applicants who received each rating level on each competency within each station. Across all stations, and competencies within station, the frequency distributions are negatively skewed. On the 1 to 7 scale, most applicants received ratings of 4, 5, or 6. Table 3 also shows the mean ratings per competency across all applicants. In general, applicants received slightly higher ratings on Station 1, and slightly lower ratings on Station 2, with the mean ratings for Stations 3 and 4 being in the middle.
<table>
<thead>
<tr>
<th>Station 1: Collaboration</th>
<th>Percentage receiving each score</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A: Collaboration</td>
<td></td>
<td>2%</td>
<td>7%</td>
<td>14%</td>
<td>21%</td>
<td>25%</td>
<td>21%</td>
<td>10%</td>
<td>4.62</td>
</tr>
<tr>
<td>1B: Maturity</td>
<td></td>
<td>2%</td>
<td>7%</td>
<td>12%</td>
<td>21%</td>
<td>24%</td>
<td>22%</td>
<td>12%</td>
<td>4.72</td>
</tr>
<tr>
<td>1C: Communication &amp;</td>
<td></td>
<td>2%</td>
<td>8%</td>
<td>13%</td>
<td>20%</td>
<td>24%</td>
<td>19%</td>
<td>13%</td>
<td>4.69</td>
</tr>
<tr>
<td>Interpersonal Skills</td>
<td></td>
<td>1%</td>
<td>5%</td>
<td>15%</td>
<td>23%</td>
<td>27%</td>
<td>20%</td>
<td>9%</td>
<td>4.65</td>
</tr>
<tr>
<td>1D: Caring</td>
<td></td>
<td>1%</td>
<td>5%</td>
<td>15%</td>
<td>23%</td>
<td>27%</td>
<td>20%</td>
<td>9%</td>
<td>4.65</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station 2: Ethical Decision Making</th>
<th>Percentage receiving each score</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A: Ethical Decision Making</td>
<td></td>
<td>4%</td>
<td>10%</td>
<td>17%</td>
<td>22%</td>
<td>20%</td>
<td>16%</td>
<td>11%</td>
<td>4.36</td>
</tr>
<tr>
<td>2B: Maturity</td>
<td></td>
<td>1%</td>
<td>9%</td>
<td>17%</td>
<td>24%</td>
<td>22%</td>
<td>17%</td>
<td>10%</td>
<td>4.48</td>
</tr>
<tr>
<td>2C: Communication &amp; Interpersonal Skills</td>
<td></td>
<td>2%</td>
<td>8%</td>
<td>18%</td>
<td>21%</td>
<td>20%</td>
<td>21%</td>
<td>10%</td>
<td>4.49</td>
</tr>
<tr>
<td>2D: Caring</td>
<td></td>
<td>2%</td>
<td>7%</td>
<td>15%</td>
<td>22%</td>
<td>29%</td>
<td>18%</td>
<td>8%</td>
<td>4.57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station 3: Reflection</th>
<th>Percentage receiving each score</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3A: Reflection</td>
<td></td>
<td>3%</td>
<td>12%</td>
<td>15%</td>
<td>18%</td>
<td>19%</td>
<td>20%</td>
<td>13%</td>
<td>4.49</td>
</tr>
<tr>
<td>3B: Maturity</td>
<td></td>
<td>2%</td>
<td>9%</td>
<td>15%</td>
<td>21%</td>
<td>22%</td>
<td>20%</td>
<td>12%</td>
<td>4.60</td>
</tr>
<tr>
<td>3C: Communication &amp; Interpersonal Skills</td>
<td></td>
<td>2%</td>
<td>9%</td>
<td>15%</td>
<td>19%</td>
<td>21%</td>
<td>20%</td>
<td>15%</td>
<td>4.67</td>
</tr>
<tr>
<td>3D: Caring</td>
<td></td>
<td>1%</td>
<td>8%</td>
<td>14%</td>
<td>25%</td>
<td>21%</td>
<td>20%</td>
<td>13%</td>
<td>4.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station 4: Values</th>
<th>Percentage receiving each score</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4A: Values</td>
<td></td>
<td>1%</td>
<td>8%</td>
<td>16%</td>
<td>19%</td>
<td>26%</td>
<td>20%</td>
<td>10%</td>
<td>4.60</td>
</tr>
<tr>
<td>4B: Maturity</td>
<td></td>
<td>2%</td>
<td>10%</td>
<td>15%</td>
<td>20%</td>
<td>21%</td>
<td>19%</td>
<td>12%</td>
<td>4.53</td>
</tr>
<tr>
<td>4C: Communication &amp; Interpersonal Skills</td>
<td></td>
<td>2%</td>
<td>8%</td>
<td>18%</td>
<td>19%</td>
<td>22%</td>
<td>18%</td>
<td>14%</td>
<td>4.59</td>
</tr>
<tr>
<td>4D: Caring</td>
<td></td>
<td>1%</td>
<td>7%</td>
<td>15%</td>
<td>20%</td>
<td>23%</td>
<td>23%</td>
<td>12%</td>
<td>4.71</td>
</tr>
</tbody>
</table>
In Table 3, from the distributions of ratings across applicants, we saw that ratings tended to be towards the higher end of the scale. It is not surprising then that, as Table 4 shows, even the most severe raters gave average scores around 2.75 to 3.75 while the most lenient raters gave average scores around 5.58 to 6.83 (recall that the maximum rating awarded is 7). Competency 1B (Maturity in Collaboration) had the highest mean rating, while Competency 2A (Ethical Decision making) had the lowest mean station rating.

**Table 4**

*Distributions of Raters’ Mean Ratings Across Stations*

<table>
<thead>
<tr>
<th>Station 1: Collaboration</th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A: Collaboration</td>
<td>3.42</td>
<td>5.92</td>
<td>4.62</td>
<td>0.57</td>
</tr>
<tr>
<td>1B: Maturity</td>
<td>3.50</td>
<td>6.17</td>
<td>4.72</td>
<td>0.56</td>
</tr>
<tr>
<td>1C: Communication &amp; Interpersonal Skills</td>
<td>3.75</td>
<td>5.92</td>
<td>4.69</td>
<td>0.51</td>
</tr>
<tr>
<td>1D: Caring</td>
<td>3.25</td>
<td>5.58</td>
<td>4.65</td>
<td>0.49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station 2: Ethical Decision Making</th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A: Ethical Decision Making</td>
<td>3.17</td>
<td>5.67</td>
<td>4.36</td>
<td>0.62</td>
</tr>
<tr>
<td>2B: Maturity</td>
<td>3.42</td>
<td>5.75</td>
<td>4.48</td>
<td>0.62</td>
</tr>
<tr>
<td>2C: Communication &amp; Interpersonal Skills</td>
<td>3.25</td>
<td>5.67</td>
<td>4.49</td>
<td>0.54</td>
</tr>
<tr>
<td>2D: Caring</td>
<td>3.33</td>
<td>5.83</td>
<td>4.57</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station 3: Reflection</th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>3A: Reflection</td>
<td>3.08</td>
<td>6.17</td>
<td>4.49</td>
<td>0.70</td>
</tr>
<tr>
<td>3B: Maturity</td>
<td>3.25</td>
<td>6.67</td>
<td>4.60</td>
<td>0.69</td>
</tr>
<tr>
<td>3C: Communication &amp; Interpersonal Skills</td>
<td>2.75</td>
<td>6.83</td>
<td>4.67</td>
<td>0.71</td>
</tr>
<tr>
<td>3D: Caring</td>
<td>3.42</td>
<td>6.67</td>
<td>4.68</td>
<td>0.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station 4: Values</th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>4A: Values</td>
<td>3.42</td>
<td>6.17</td>
<td>4.60</td>
<td>0.66</td>
</tr>
<tr>
<td>4B: Maturity</td>
<td>3.33</td>
<td>6.50</td>
<td>4.53</td>
<td>0.69</td>
</tr>
<tr>
<td>4C: Communication &amp; Interpersonal Skills</td>
<td>3.42</td>
<td>6.50</td>
<td>4.59</td>
<td>0.67</td>
</tr>
<tr>
<td>4D: Caring</td>
<td>3.55</td>
<td>6.17</td>
<td>4.71</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Table 5 shows the correlations among competency ratings. The correlations are meant to provide an indicator of inter-rater agreement of commonly rated competencies. What is surprising is that the highest correlations among competency ratings are within stations; these are highlighted in grey ($r = .70-.86$). Although, raters at every station rated three common competencies, the correlations among those competencies across stations are surprisingly low ($r = .06-.27$). For example, the correlation between 1B and 3B is .13, which is quite low, considering both were assessing maturity at different stations.

Table 5

<table>
<thead>
<tr>
<th></th>
<th>1A</th>
<th>1B</th>
<th>1C</th>
<th>1D</th>
<th>2A</th>
<th>2B</th>
<th>2C</th>
<th>2D</th>
<th>3A</th>
<th>3B</th>
<th>3C</th>
<th>3D</th>
<th>4A</th>
<th>4B</th>
<th>4C</th>
<th>4D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B</td>
<td>0.79</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1C</td>
<td>0.77</td>
<td>0.77</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1D</td>
<td>0.70</td>
<td>0.73</td>
<td>0.71</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2A</td>
<td>0.18</td>
<td>0.16</td>
<td>0.16</td>
<td>0.14</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2B</td>
<td>0.15</td>
<td>0.17</td>
<td>0.16</td>
<td>0.12</td>
<td>0.85</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2C</td>
<td>0.23</td>
<td>0.21</td>
<td>0.24</td>
<td>0.19</td>
<td>0.79</td>
<td>0.85</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
<td>0.15</td>
<td>0.77</td>
<td>0.81</td>
<td>0.79</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3A</td>
<td>0.12</td>
<td>0.12</td>
<td>0.22</td>
<td>0.17</td>
<td>0.17</td>
<td>0.20</td>
<td>0.23</td>
<td>0.15</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3B</td>
<td>0.12</td>
<td>0.13</td>
<td>0.23</td>
<td>0.16</td>
<td>0.16</td>
<td>0.21</td>
<td>0.23</td>
<td>0.17</td>
<td>0.86</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3C</td>
<td>0.11</td>
<td>0.12</td>
<td>0.23</td>
<td>0.15</td>
<td>0.17</td>
<td>0.21</td>
<td>0.27</td>
<td>0.19</td>
<td>0.83</td>
<td>0.85</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3D</td>
<td>0.08</td>
<td>0.06</td>
<td>0.19</td>
<td>0.14</td>
<td>0.14</td>
<td>0.16</td>
<td>0.21</td>
<td>0.16</td>
<td>0.81</td>
<td>0.80</td>
<td>0.79</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4A</td>
<td>0.11</td>
<td>0.14</td>
<td>0.19</td>
<td>0.17</td>
<td>0.15</td>
<td>0.18</td>
<td>0.21</td>
<td>0.17</td>
<td>0.19</td>
<td>0.18</td>
<td>0.21</td>
<td>0.15</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4B</td>
<td>0.15</td>
<td>0.17</td>
<td>0.23</td>
<td>0.17</td>
<td>0.15</td>
<td>0.17</td>
<td>0.21</td>
<td>0.16</td>
<td>0.20</td>
<td>0.19</td>
<td>0.22</td>
<td>0.14</td>
<td>0.84</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4C</td>
<td>0.15</td>
<td>0.18</td>
<td>0.25</td>
<td>0.20</td>
<td>0.15</td>
<td>0.17</td>
<td>0.24</td>
<td>0.20</td>
<td>0.18</td>
<td>0.17</td>
<td>0.24</td>
<td>0.15</td>
<td>0.79</td>
<td>0.80</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>4D</td>
<td>0.11</td>
<td>0.13</td>
<td>0.20</td>
<td>0.17</td>
<td>0.12</td>
<td>0.14</td>
<td>0.15</td>
<td>0.14</td>
<td>0.15</td>
<td>0.14</td>
<td>0.16</td>
<td>0.14</td>
<td>0.84</td>
<td>0.79</td>
<td>0.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Research Question 2: How well is the rating behaviour modelled by the MFRM? How lenient and severe are the raters?

The results will explore whether observed scores (received by applicants or awarded by raters) reasonably fit the expected scores calculated by the model. The parameters estimated in the analyses are outlined below.

Applicant Ability (theta-A): This is positively oriented within the model, which means that along the x-axis, the more positive the theta-A value, the higher the applicant ability.

Rater Severity (theta-R): This is negatively oriented within the model, which means that along the x axis, the positive the theta-R value, the more systematically severe the rater was in their scoring.

Figure 1 shows a scatterplot of the applicants’ mean rating versus their ability level. The y-axis represents the rating scale (from 1 to 7) applicants were scored on. The x-axis represents the ability level of applicants which is the theta-A value calculated for each applicant by the MFRM model. The distribution of abilities was centered on zero with a standard deviation of 1. Not surprisingly, we see a positive relationship between ability level and average ratings awarded. The blue points in the figure are the actual mean ratings and the red points are the estimated (fair) mean ratings derived from the MFRM. The fair mean ratings represent a calculated estimate of the mean ratings that corresponds with the applicants’ ability level. In effect, the fair mean ratings are the ratings the applicant would have received if not for variations in raters’ severity. Though both the fair and actual mean ratings show a positive relationship between the ratings awarded and ability level, what becomes apparent with this comparison is the amount of y-axis variation seen in the actual mean ratings. To explore whether this variation
can be considered meaningful or would warrant concern, we turn to the infit mean-square analysis presented below.

![Figure 1. Applicants’ mean ratings against theta-A.](image)

Figure 2 shows the applicants’ infit mean square indices against applicants’ ability (theta-A) generated by the MFRM. The area between the dotted lines are the generally accepted infit range (0.75-1.3). Infit mean square indices that are above 1.3 demonstrate misfit of the model (the observed ratings are farther away than what the model would have predicted); 23% of the applicants have an infit above 1.3. Thus, further investigation into how rater effects systematically contribute to this variation is needed. The following sections will present analogous plots of mean ratings and infit indices against rater severity (theta-R) derived by the MFRM.
Figure 2. Applicants’ fit mean squares against applicants’ ability estimates (theta-A).

Figure 3 shows the mean ratings of each rater against rater severity (theta-R). Rater severity was negatively oriented within the model; therefore, lower (or more negative values) represent rater leniency and higher (more positive values) represent rater severity. Raters from different stations are represented by different symbols (station numbers correspond with station-specific competencies). As we can see, severe raters, on average, awarded lower scores while lenient raters, on average, awarded higher scores. This pattern is seen across all stations. The stars in Figure 3 are the fair mean ratings derived from the MFRM, which are a calculated estimate of the mean ratings that the model predicts raters would award depending on the rater severity indices (Theta-R values) derived by the model. The fair mean ratings are across all
stations, and we can see that both the fair and actual station mean ratings are well aligned. To check whether this fit is statistically meaningful, we turn to the rater infit mean-square analysis presented below.

Figure 3. Raters’ mean ratings against theta-R.

Figure 4 shows the rater infit mean square indices against rater severity (theta-R) generated by the MFRM analysis. The area between the dotted lines is the generally accepted infit range (0.75-1.3). Infit mean square indices that are above 1.3 demonstrate misfit (the observed ratings are farther away than what the model would have predicted); 21% of the raters are above 1.3.
In Figures 1 to 4, applicants’ and raters’ estimates were considered separately. Figure 5 displays the Wright Map (variable map) which puts raters, applicants, stations, and scores onto a vertical logit ruler. The facets are seen within a single frame of reference; this allows for visual comparisons of the variation within and across each facet. The first column (from left to right) represents the logit ruler which is the scale that all of the facets in the analysis are placed along. The logit ruler can range from – infinity to + infinity, but shown here, the range is from -3 to 3. The last column represents the rating scale raters used during the interview process (ranging from 1 to 7). The second column represents the rater facet, and depicts variation in rater severity.

Figure 4. Applicants’ infit mean squares against applicant ability (theta-R).
The rater facet was negatively oriented within the MFRM; therefore, for interpretation of this column, raters who are more positive on the logit ruler are systematically more severe in their rating behavior. The third and fourth columns represent the applicant facet and the station facet, both of which were positively oriented in the model. The applicant column depicts variation in scores due to applicant ability, and the station column depicts variation in station score due to differences in station difficulty. For interpretation, applicants who are more positive on the logit ruler are high ability applicants who (looking at the last column) were systematically awarded higher ratings. The station column represents the difficulty level of each station. For interpretation, the higher the station, the less difficult it was for applicants.
Figure 5. Wright Map of raters, applicants, stations, and MPI scores.
Research Question 3: How are modes of training related to rater behaviour?

As Table 6 shows, the majority of raters used the online training. About a quarter (25.5%) of the raters did both online and in-person training, 63.0% did online training only, and 3.5% did in-person training only. Seven raters (3.5%) did neither type of training. However, all raters did attend the on-site training immediately before the interviews.

Table 6

Raters’ Participation in Training

<table>
<thead>
<tr>
<th>Viewed One or More Online Training Module</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attended In-person Training</td>
<td>16</td>
<td>126</td>
<td>142</td>
</tr>
<tr>
<td>No</td>
<td>16 (8.0%)</td>
<td>126 (63.0%)</td>
<td>142 (71.0%)</td>
</tr>
<tr>
<td>Yes</td>
<td>7 (3.5%)</td>
<td>51 (25.5%)</td>
<td>58 (29.0%)</td>
</tr>
<tr>
<td>Total</td>
<td>23 (11.5%)</td>
<td>177 (88.5%)</td>
<td>200 (100.0%)</td>
</tr>
</tbody>
</table>

Table 7 shows training participation by station. The mode(s) of training participated in by raters were not significantly related to the station to which the rater was assigned ($\chi^2 (9) = 7.87$, $p = .55$). That is, there was no significant difference in training mode participation across stations.
Table 7

Raters’ Participation in Training by Station

<table>
<thead>
<tr>
<th>Station</th>
<th>No Training</th>
<th>Online Only</th>
<th>In-person Only</th>
<th>Both In-person and On-line</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Collaboration</td>
<td>2</td>
<td>32</td>
<td>0</td>
<td>16</td>
<td>50</td>
</tr>
<tr>
<td>2: Ethical Decision Making</td>
<td>4</td>
<td>33</td>
<td>1</td>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>3: Reflection</td>
<td>6</td>
<td>28</td>
<td>3</td>
<td>13</td>
<td>50</td>
</tr>
<tr>
<td>4: Values</td>
<td>4</td>
<td>33</td>
<td>3</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>All Stations</td>
<td>16</td>
<td>126</td>
<td>7</td>
<td>51</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 8 shows the average ratings (Theta-R) by station and training mode. A two-way analysis of variance found no significant differences among stations ($F(3,185) = 0.08, p = .97$, partial $\eta^2 = 0.001$) or among training modes ($F(3,185) = 0.57, p = .64$, partial $\eta^2 = 0.01$) or the interaction ($F(8,185) = 0.58, p = .79$, partial $\eta^2 = 0.03$). A three-way analysis of variance in which online training and in-person training were treated as separate factors was also not significant.
### Table 8

**Average Theta for Raters who Participated in Different Modes of Training by Station**

<table>
<thead>
<tr>
<th>Station</th>
<th>No Training (N = 16)</th>
<th>Online Only (N = 126)</th>
<th>In-person Only (N = 7)</th>
<th>Both In-person and On-line (N = 51)</th>
<th>Total (N = 200)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Collaboration</td>
<td>-0.57 (0.62)</td>
<td>-0.55 (0.36)</td>
<td>n/a</td>
<td>-0.53 (0.29)</td>
<td>-0.54 (0.34)</td>
</tr>
<tr>
<td>2: Ethical Decision Making</td>
<td>-0.60 (0.40)</td>
<td>-0.51 (0.35)</td>
<td>-0.42 (0.00)</td>
<td>-0.38 (0.44)</td>
<td>-0.48 (0.37)</td>
</tr>
<tr>
<td>3: Reflection</td>
<td>-0.67 (0.33)</td>
<td>-0.49 (0.55)</td>
<td>-0.49 (0.71)</td>
<td>-0.51 (0.43)</td>
<td>-0.52 (0.50)</td>
</tr>
<tr>
<td>4: Values</td>
<td>-0.67 (0.35)</td>
<td>-0.44 (0.39)</td>
<td>-0.34 (0.25)</td>
<td>-0.74 (0.70)</td>
<td>-0.51 (0.47)</td>
</tr>
<tr>
<td>All Stations</td>
<td>-0.64 (0.35)</td>
<td>-0.50 (0.41)</td>
<td>-0.57 (0.38)</td>
<td>-0.53 (0.46)</td>
<td>-0.51 (0.42)</td>
</tr>
</tbody>
</table>

Table 9 summarizes the raters’ inﬁt by station and training mode. A two-way analysis of variance found no significant differences among stations ($F(3,185) = 0.471, p = .70,$ partial $\eta^2 = 0.008$) or among training modes ($F(3,185) = 1.24, p = .30,$ partial $\eta^2 = 0.020$) or the interaction between station and training mode ($F(8,185) = 0.95, p = .47$ partial $\eta^2 = 0.040$).
Table 9

*Average Absolute Infit for Raters who Participated in Different Modes of Training by Station*

<table>
<thead>
<tr>
<th>Station</th>
<th>No Training ( (N = 16) )</th>
<th>Online Only ( (N = 126) )</th>
<th>In-person Only ( (N = 7) )</th>
<th>Both In-person and On-line ( (N = 51) )</th>
<th>Total ( (N = 200) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Collaboration</td>
<td>0.35 (0.07)</td>
<td>1.78 (1.32)</td>
<td>(n/a)</td>
<td>1.23 (1.05)</td>
<td>1.54 (1.25)</td>
</tr>
<tr>
<td>2: Ethical Decision Making</td>
<td>1.65 (1.20)</td>
<td>1.59 (1.52)</td>
<td>0.60 (0.00)</td>
<td>2.10 (1.35)</td>
<td>1.70 (1.44)</td>
</tr>
<tr>
<td>3: Reflection</td>
<td>1.20 (0.85)</td>
<td>1.48 (1.00)</td>
<td>0.83 (1.01)</td>
<td>1.26 (0.98)</td>
<td>1.35 (0.97)</td>
</tr>
<tr>
<td>4: Values</td>
<td>1.23 (0.88)</td>
<td>1.59 (1.28)</td>
<td>1.77 (1.53)</td>
<td>0.93 (0.66)</td>
<td>1.44 (1.17)</td>
</tr>
<tr>
<td>All Stations</td>
<td>1.21 (0.91)</td>
<td>1.61 (1.29)</td>
<td>1.20 (1.18)</td>
<td>1.38 (1.11)</td>
<td>1.51 (1.22)</td>
</tr>
</tbody>
</table>
CHAPTER 5. DISCUSSION AND CONCLUSION

The purpose of this study was two-fold: to examine rater effects through both descriptive and psychometric analyses of interview scores of applicants to the University of Toronto’s Undergraduate Medicine Program, and see whether different modes of rater training were related to rater scoring behaviours. The interview format was the modified personal interview, developed at the University of Toronto, which blends multiple independent sampling methods and personal interview formats. Applicants rotated through four stations in which they were evaluated on four main competencies within stations (collaboration, ethical decision making, reflection, and values) and three across-station competencies (maturity, communication and interpersonal skills, and caring). Raters rated applicants on a 7-point Likert-type scale (1 being unsuitable and 7 being superb). Along with interview data, rater training data were also analyzed. Raters had the option of attending online training and/or in-person training; all raters were required to attend day of training.

In the first part of the analyses, the frequency distributions of the station mean scores of each competency demonstrate negative skewness; raters, on average, awarded applicants high scores. Station 1 received slightly higher ratings, Station 2 had slightly lower ratings, with Stations 3 and 4 in the middle. These results are also echoed within the Wright Map from the MFRM analyses: Station 2 was systematically more difficult than Stations 1, 3, and 4. Stations 1 (collaboration), 3 (reflection), and 4 (values) require applicants to provide introspective responses regarding past behaviours and experiences or to reflect on a statement. In Station 2 (ethical decision making), the applicants were presented situational-judgement questions based on dilemmas currently facing the healthcare system. To arrive at a response, the applicant would
need not only to give an answer, but also to demonstrate that they can see the problem with multiple frames of reference, as well as consider the consequences of their decisions from different perspectives. It is not surprising, then, that this station would be more difficult in comparison to the others.

The frequency distributions of the rater means also reflect these findings. The distributions of the rater means show noticeable rater variation; the most severe raters had means around 2.75 to 3.75 while the most lenient raters had means around 5.58 to 6.83. Even though the rating scale has 7 rating categories, raters are concentrating their rating judgements within the highest part of the rating scale.

The correlations of ratings across competencies also provided insight into the competencies that were intended to be the same across stations. As stated earlier, a goal of the MPI format was to provide reliable measurement, while maintaining the personal interview to maximize recruitment potential. Assessing common competencies across stations (maturity, communication and interpersonal skill, and caring) was intended to provide multiple independent ratings to improve measurement of those competencies. The more times an applicant is independently assessed on a competency, the more reliable the measure. Therefore, what is expected is that there would be higher rating correlations among common competencies across stations. Instead, what is observed is markedly low correlations for the common competencies across stations ($r = .06-.27$), and higher correlations across competencies within station ($r = .70-.86$). This suggests that context specificity is still affecting the ratings within the MPI format. The station could have a large influence on how applicants were rated. A second contributor to the high within-station correlations could be halo effect. As each station only had one rater, it is
quite possible that overall global impressions of the applicants may have influenced the raters’ scoring patterns for all competencies, instead of them rating each competency in isolation.

The second part of the analysis employed MFRM to examine how rater effects (in particular rater severity and rater leniency) affect the applicants’ interview scores. Scatterplots of the applicants’ estimated ability levels against the mean scores applicants received showed trends that were expected; higher ability applicants on average received higher mean interview scores. However, an index of model fit suggested that many of the applicants received ratings that were more variable than expected. Because these analyses were performed across all competencies and stations, this lack of fit may suggest that the 16 ratings received by each applicant are measuring more than a single underlying ability. The low correlations across the stations may also suggest that the MPI is measuring more than one ability, although these correlations could also be due to context effects or halo effects.

The MFRM also examined the rater facet to see how differences in rater effects could have contributed to the variation seen in applicant ratings. Many of the raters also did not fit the model well, suggesting that some raters were not consistently lenient or severe in relation to other raters.

The most telling piece of output produced by the MFRM analysis was the Wright Map (variable map). Placing all of the facets along a logit ruler (for these purposes, ranging from -3 to 3), the Wright map allowed comparison of the variability within and between facets (raters, applicants, and stations). Within the rater facet, the variation of rater severity spread across approximately 2 logits, while the spread of applicant ability spanned 5.5 logits. In the Wright Map, the estimates for the applicants control for the differences in rater severity (similarly, the differences in rater severity control for the differences in the applicants). Figure 1 showed that
there is some reordering of the applicants when scores are adjusted for differences in rater severity. If a goal of the interview is to discriminate between high ability and low ability applicants, then these findings bring to light the importance of psychometric analyses in making visible what other facets may affect ratings. This kind of analysis also has implications for fairness in applicant assessment; a future direction would be to see how rater characteristics (for example, professional or academic backgrounds of raters) are linked to their rater severity measures. As well, exploring how demographic variables of applicants relate to their estimated abilities and observed ratings may also be interesting and have implications for access.

The third part of the study examined whether different modes of rater training were related to raters’ behaviours. The majority of raters favoured the online-training, with around 25% of raters participating in both the online and in-person training. A chi-square test performed showed that there was no significant difference in the type of training mode participation across stations. Analysis of variance of severity estimates and fit indices for raters across station and training found no significant differences. However, that almost 90% of the raters participated in the online training and fewer than 10% did no training makes it difficult to make conclusions about the effects of training. Adding to the difficulty is the lack of information about whether raters took part in training in previous years.

Limitations of the Study and Future Directions

Because raters were nested within station (that is, each rater was at only one station) and sets of four raters rated the same 12 applicants, the data consisted of disjointed subsets, which meant that the MFRM analysis could only partially adjust for differences in severity across raters and in differences in ability across applicants. A recommendation for future data collection is to
explore the possibility of having raters change station during the interview process and also have the replacement of raters staggered instead of replacing all four raters after every 12 applicants.

This study did not include a formal investigation of the dimensionality of the ratings. Because the results suggest possible multidimensionality, such an investigation would be important for future analyses.
REFERENCES


http://doi.org/10.1046/j.1365-2923.2003.01563.x


http://doi.org/10.1023/B:AHSE.0000027479.14637.6c


http://doi.org/10.1046/j.1365-2923.2004.01776.x


http://doi.org/10.1002/9781444320282.ch24


systematic review. *Medical Teacher, 35*(12), 1027–1041.

http://doi.org/10.3109/0142159X.2013.829912


http://doi.org/10.1023/A:1011489618208


http://doi.org/10.1097/ACM.0000000000000902


