Canadian Journal of Counselling and Psychotherapy /
Revue canadienne de counseling et de psychothérapie
ISSN 0826-3893 Vol. 48 No. 2 © 2014 Pages 108–126

Assessment of a Master of Education Counselling Application Selection Process Using Rasch Analysis and Generalizability Theory
Évaluation d'un processus de sélection de candidats pour un programme de maîtrise en éducation en counseling en utilisant l'analyse Rasch et la théorie de la généralisabilité

Stefanie S. Sebok
Queen's University
Peter D. MacMillan
University of Northern British Columbia

ABSTRACT

This study was designed to evaluate an application selection process for a Master of Education counselling program in Canada using the Many-Facet Rasch Model (MFRM) and Generalizability Theory (G-Theory). Current literature pertaining specifically to counselling admissions is essentially absent. This study investigated the items used to score and rank applicants as well as rater characteristics for each of the members of the application selection committee. The design, results, and findings have implications for admissions procedures and practices at other universities within Canada. Overall, the MFRM and G-Theory functioned as appropriate measurement tools for assessing counselling admission items, raters, and applicants.

RÉSUMÉ

Cette étude a évalué l'efficacité du modèle de mesure à multi-facettes de Rasch (MFRM) et de la théorie de la généralisabilité lorsqu'appliqués à un processus de sélection d'applications pour un programme de maîtrise d'éducation en counseling au Canada. Il existe présentement un vide quant à la recherche se rapportant spécifiquement aux admissions dans des programmes de counseling. Ainsi, cette étude a examiné les éléments utilisés pour évaluer et classifier chaque demande, en plus de considérer les caractéristiques des évaluateurs pour chacun des membres du comité de sélection. La conception, les résultats, et les conclusions ont des implications pour les procédures et les pratiques d'admission dans d'autres universités au Canada. Dans l'ensemble, le MFRM et la théorie de la généralisabilité fonctionnaient comme des outils de mesure appropriés afin d'évaluer les éléments rattachés aux admissions dans un programme de counseling, soit les items, les évaluateurs et les candidats.

Departments and faculties in universities everywhere are faced with the challenge of deciding, usually on an annual or semiannual basis, which individuals should be offered the opportunity to study at their particular institution. The application selection process has become even more challenging for highly competitive graduate and professional programs where there are a large number of individuals applying for a limited number of available spaces. Given that the pressure to select the best possible candidates without bias continues to grow, graduate and professional schools are becoming increasingly interested in evaluating the effectiveness of their own admissions processes to help ensure that their admissions processes are fair and accurate. Furthermore, not only do universities want selection methods that offer admission to those prospective students who have the desired qualities and characteristics needed for success in a particular program, but they also want to deny admission to applicants who may be problematic (Homrich, 2009). These dual purposes for admissions procedures exist because, in many graduate and professional programs, failure of candidates is not an option.

University counselling programs provide one example of a professional program where the admission decisions are "high stakes." In Canada, over 40 universities have master's-level programs in counselling and admit students annually (Canadian Counselling and Psychotherapy Association, 2013). At most institutions, a large number of quality applicants vie for a limited number of spaces; therefore, selecting individuals who have strong interpersonal skills and are likely to complete the program within a timely fashion is challenging because the decisions are often based on fine distinctions. Due to the high-stakes competitive nature of graduate programs, many applicants likely employ the strategy of applying to multiple institutions. However, some applicants may receive multiple rejections as they are not ranked favourably according to any university's admissions standards. Given that a master's degree is typically required to become certified as a counsellor, an inability to secure an admissions offer ends, at least temporarily, the applicant's chances of a career in counselling. The possibility of never working in a career that one had hoped and prepared for makes the application process high-stakes. Typically, the preadmissions criteria used to screen potential counselling students are previous academic performance, writing ability, goals of the applicant, professional experience, and letters of reference (Corey, Corey, & Callanan, 2007; Nelson, Canada, & Lancaster, 2003).

Unlike many other graduate programs that select applicants based solely on academic performance and research potential, the counselling application process includes a combination of academic and nonacademic admissions criteria in the selection process, which at times can be difficult to accurately measure. Furthermore, given the responsibilities of the profession, there is a need to ensure that those students entering counselling programs, and subsequently the counselling profession, have a specific set of qualities and characteristics. Previous research examining the selection criteria for counselling admissions in Canadian contexts has largely focused on doctoral programs in counselling (Pass & Scherer, 1979). Current research investigating the admissions process of master's-level counselling programs within Canadian institutions is essentially nonexistent, which is problematic given that within Canada only a master's degree in an accredited counselling program is required to become certified as a counsellor. Therefore, admissions decisions made at the master's level are a crucial form of gatekeeping because, once an applicant enters the program, the impetus is to keep the individual in the program, using mechanisms to remediate any deficiencies. Instances where programs have high proportions of individuals requiring remediation are problematic because time and resources are redirected toward supporting those requiring remediation.

Counselling admissions processes, like many other admissions processes, rely specifically on trained professionals with experience and/or expertise in a specific area to aid in decision-making. The issue with using human beings in application selection is the subjectivity involved in the process. When people make judgements there is often bias, whether identified or not, that influences how each individual views and interprets the suitability and quality of an applicant. Therefore, for an admissions process that is accurate and reliable, utilization of resources that can help minimize high levels of variation and inconsistency caused by subjectivity is crucial. Many objective forms of measurement exist that can be used to support decisions made by an application selection committee. Statistical information about persons (i.e., the applicants in this study), raters, and items—all known to be common sources of variation—are obtained using measurement models. Such information is necessary to determine the extent in which persons, raters, and items differ, as the differences are often substantial. Further, knowledge about each preadmission item's level of difficulty is worthwhile, to ensure that the measures used to screen counselling students are neither too easy or too difficult. When items used for admissions purposes are not appropriately matched to the ability levels of the applicants, the chances of selecting unsuitable candidates increase. Through the use of various measurement models, adjustments can be made to control and correct for differences and inconsistencies found in the admissions process (Linacre, Wright, & Lunz, 1990).

The purpose of our research was to assess the overall effectiveness of the Master of Education counselling (MEDC) applicant selection process as it currently exists at one Canadian university. We approached assessment of the applicant selection process from a measurement perspective, using the Many-Facet Rasch Model (MFRM) and Generalizability Theory (G-Theory). Through an analysis of the items used to score applicants as well as the rater characteristics of each of the members on the application selection committee, we focused on the following subquestions:

1. What are the characteristics of the items used to assess MEDC applicants, and are item difficulties appropriately matched to applicants' ability levels?
2. What is the rating behaviour of the faculty and students who participated in selecting applicants for the MEDC program both as individuals and as a group?

The characteristics of the applicants (i.e., application files) are not explicitly discussed in this article, given that they are subsumed in the rater and item analyses.

THE MANY-FACET RASCH MODEL

The Rasch model was designed to objectively analyze data by creating a scale that measures it consistently across a population. The Danish mathematician Georg Rasch created the original Rasch model in 1965, based on the principles of log-odds transformations and additive measurement (Stone, Beltyukova, & Fox, 2008). He used this model to examine person ability and item difficulty in dichotomously scored data (Rasch, 1980). The foundational element of the Rasch model was to identify persons and items that are not performing as expected. Since then, others have advanced the concept of Rasch modelling to create the Rating Scale Model (Andrich, 1978), the Partial Credit Model (Masters, 1982), and the MFRM (Linacre, 1989).

The MFRM goes beyond person ability and item difficulty to measure other factors (severity of judges, differences in rating scales, and consistency across occasions) that can interact within a testing situation (e.g., Chang & Chan, 1995; MacMillan, 2000). Hence, the MFRM is an effective measure of observed rater effects (e.g., Engelhard, 1994; Kim & Wilson, 2010; Linacre et al., 1990; O'Neill, 1999; Wolfe, 2004). Rater effects are any systematic patterns of unconventional behaviour that exist within an individual's rating practices (Wolfe, 2004). The most common forms of rater effects include halo, severity or leniency, and central tendency (Eckes, 2005). A halo effect can occur in one of two ways: when a previous rater influences ratings of a subsequent rater (Linacre, 2010) or when a rater assesses a person holistically rather than on an item-by-item basis (Engelhard, 1994). Severity is used to describe a rater who consistently rates below the midpoint of a scale, while leniency is when a rater generally awards scores above the midpoint (Myford & Wolfe, 2004). Central tendency results when a rater overuses the middle categories, unintentionally avoiding the outermost categories (Myford & Wolfe, 2004). These rater effects create variability that results in different decisions based on individual raters. Rater effects result in biased assessments, which lead to inaccurate applicant rankings. In the case of the counselling program, we used the MFRM to explore any rater effects that exist in the assessment of MEDC applicants.

GENERALIZABILITY THEORY

Cronbach, Glaser, Nanda, and Rajaratnam (1972) first introduced the concepts of G-Theory by extending the work done by Hoyt in 1941 (Kieffer, 1999). Using some of the same principles of traditional ANOVA, G-Theory uses variance components to represent the amount of error that comes from generalizing from a facets score to a universal score (Swiss Society for Research in Education Working Group, 2010). In any measurement situation, there is a desire to obtain scores that are able to accurately separate the performance of different examinees while also minimizing the variability in the other factors (e.g., items or raters). Variability in these other factors (facets) reduces the accuracy in the measurement of examinee performance. G-Theory examines the extent to which each facet individually contributes to variation (error) in the measurement of a person's overall score in order to obtain a better account of the person's true ability, and thus makes inferences that can be generalized back to the population. G-Theory measures the impact of these facets individually and among the interactions between each facet (Shavelson & Webb, 1991).

Over the last decade countless studies have assessed and analyzed persons, raters, items, and occasions using G-Theory (e.g., Harik et al., 2009; Pedersen, Hagtvet, & Karterud, 2007; Smith & Kulikowich, 2004; Winne et al., 2006). More specifically, G-Theory has been applied to identify sources of variance in examination processes used to select applicants for specialized university programs (Atilgan, 2008; Oosterveld & ten Cate, 2004). Admissions processes, clinical assessments, and licensure examinations are all high-stakes situations that strive to have standardized measures to avoid unfair advantages or disadvantages among individuals. Therefore, the need to be unbiased and consistent in high-stakes situations has prompted investigations about the generalizability of various facets (i.e., persons, raters, and items). Based on traditional ANOVA methods, G-Theory can separately evaluate multiple sources of measurement variance (Atilgan, 2008). Furthermore, by using G-Theory to examine all of the identified main effects and interactions individually, we can account for the unexplained sources of variability and produce a G-coefficient that reflects the true amount of variance associated with a person's score (Shavelson & Webb, 1991).

METHODS

Participants

Raters. The participants for our study were 3 faculty members and 2 graduate students who reviewed the applicants for a graduate counselling program and determined which applicants would be offered admission into the program. The 3 faculty members were from the School of Education and taught in the counselling program. Two of the 3 faculty members were certified counsellors, while the third is a measurement specialist who taught statistics and supervised counselling research for over a decade. One faculty counsellor is now a tenured associate professor, while the other is a term assistant professor. Both have over 20 years of clinical counselling experience. The 2 graduate students were both near completion of their counselling degrees and graduated before the successful applicants entered the program. Both students performed in the top 5% of their peer group, worked as certified counsellors, and are presently engaged in doctoral studies.

Applicants. The applicant pool consisted of applicants to the MEDC program. The population applying was roughly 80% female and 20% male. Applicants ranged in age from 22 to 55. Most applicants had previously obtained a bachelor's degree in psychology, social work, criminology, or education. Finally, their levels of relevant work experience ranged from those with some volunteer experience in a helping arena to those who had been employed in the counselling field for over 30 years. Approximately 20–35% of applicants are given a letter of offer in any given year. The size of the program is limited by the availability of suitable practicum placement opportunities in the small-medium-sized Canadian city where the university is based.

Measures

Each application package consisted of (a) a grade-point-average (GPA), (b) relevant degree information, (c) written evidence of involvement with people in appropriate settings, (d) a written personal statement, and (e) three letters of reference. These preadmissions criteria were developed by faculty members in the School of Education and were consistent with measures used in previous years to select counselling students. The 3 faculty members and the 2 students rated every MEDC applicant on all of the preadmissions criteria, with the exception of GPA, using a series of 5-point scales. Applicants were rated on 10 items: relevant educational degree, writing ability, fit of personal goals with the MEDC program, work experience, suitability of first referee, quality of the applicant according to the first reference letter, suitability of second referee, quality of the second reference letter, suitability of third referee, and quality of the third reference letter. All 5-point scales were similar in that a low value indicated less desirable performance and a high value indicated more desirable performance. For example, the scale used to assess competency of written communication consisted of 1 = *very poorly written*, 2 = *poorly written*, 3 = *acceptable*, 4 = *well written*, and 5 = *very well written*. An overall score based on all of the application criteria was used to rank the applicants. The rank-ordering information generated throughout this study was used to make final decisions about who would be offered a seat in the program.

Procedure

This study was reviewed and supported by the University's Research Ethics Board. All 3 faculty members and 2 graduate students who reviewed the applications signed consent forms agreeing to participate in this research. Following the university's established procedure, all applications were initially collected by the registrar's office. Applicant packages of individuals who met the GPA requirement (as well as those individuals who did not meet the GPA requirement, but were specifically requested by the counselling coordinator) were forwarded to the Chair of the application selection committee, who checked them over and prepared them for the committee. In order to ensure anonymity of applicants, all application packages were coded for the participants by the Chair. The Chair had these applications photocopied, with all identifiers removed, for each of the committee members. A rater training session was developed and delivered by the first author, under the direction of the second author who possessed professional experience in rater training and protocol. The 3 faculty and 2 student raters all participated in a 3-hour rater training session, which described the established selection process and reviewed the 5-point scales used for rating applicants. Rater training also included a discussion about some of the most common rating effects (halo, severity or leniency, and central tendency) that have been shown to be problematic. Following the training session, all raters were given copies of the application packages for the 49 candidates applying to the MEDC program. All applications were read and scored within a two-week period as agreed upon by the application selection committee. Once all 5 raters finished scoring every application package, the packages were returned to the Chair of the selection committee and the raters met for an hour debrief and follow-up.

Data Analysis

GPA and all other ratings for each applicant were entered into an EXCEL file. The data obtained from the raters were analyzed using FACETS, version 3.03 (Linacre, 1996) and EDUG, version 6.0 (Swiss Society for Research in Education Working Group, 2010).

Rasch analysis. The research design was a fully crossed three-facet MFRM, examining applicants' ability (based on the quality of their counselling application), the difficulty of the items on the 5-point scale, and the severity of all the raters on the application selection committee. The Rasch-Andrich Rating Scale, described in Linacre (2010), was employed. This measurement model was used for the analysis because it allows for examination of interactions between multiple facets. Each facet was examined to see the level of influence it had on the probability of a particular applicant scoring the way they did on specific items by various raters.

Generalizability analysis. In addition to the Rasch analysis, we conducted a fully crossed two-facet (item and rater) generalizability analysis. The G-Theory analysis enabled us to examine the different sources of variability that existed within the admissions process (e.g., Shavelson & Webb, 1991).

Results

Many-Facet Rasch Analysis

The results of the many-facet Rasch analysis are best shown using a Wright Map (Figure 1). The far left column (*Measr*) is the logit scale used to measure all of the facets within the design. The second column (+*Student*) is the distribution of the applicants; most of the applicants were situated within the 0 to 2 region on the logit scale, indicating that the applicants were judged proficient. The third column (-*Program*) contains the program status: full-time or part-time studies, which in Figure 1 demonstrates that the full-time and part-time applicant pools are of equal ability given their location on the logit scale. The fourth column (-*Rater*) is the rater facet. Notice that all of the raters were positioned around the 0 logit mark. Those raters above 0 logits would be considered more severe (see R1 and R3), while those below 0 logits would be considered less severe (see R4 and R5). These differences, although visible, are relatively small. The fifth column (-*Items*) represents the item difficulties; more difficult items are in the positive logit region (e.g., item 4: work experience) and the less difficult items are in the negative logit

Figure 1.
Wright variable map for relationships among facets for counselling applicants

```
|Measr|+Student|-Program     |-Rater   |-Items   |S.1  |
+   3 +        +             +         +         +(5)  +
|     |        |             |         |         |     |
|     |        |             |         |         |     |
|     | *      |             |         |         |     |
|     |        |             |         |         |     |
|     |        |             |         |         |     |
|     |        |             |         |         |     |
|     |        |             |         |         |     |
|     | **     |             |         |         | --- |
|     |        |             |         |         |     |
+   2 +        +             +         +         +     +
|     |        |             |         |         |     |
|     | *      |             |         |         |     |
|     | *      |             |         |         |     |
|     | ****   |             |         |         |     |
|     | ***    |             |         |         |     |
|     | ****   |             |         |         |     |
|     | *****  |             |         |         |     |
|     | *****  |             |         |         | 4   |
|     | ****   |             |         |         |     |
+   1 + ****   +             +         +         +     +
|     | *      |             |         |         |     |
|     | ****   |             |         | 4       |     |
|     |        |             |         |         |     |
|     | ****   |             |         |         |     |
|     | *      |             |         |         |     |
|     | **     |             |         | 8       | --- |
|     |        |             |         | 1    10 |     |
|     |        |             |         | 2       |     |
|     | *      |             | R1   R3 | 6       |     |
*   0 * *      *  FT     PT  * R2      *         *     *
|     |        |             | R4      |         |     |
|     |        |             | R5      | 3       | 3   |
|     | *      |             |         |         |     |
|     |        |             |         | 9       |     |
|     |        |             |         |         |     |
|     |        |             |         | 7       |     |
|     |        |             |         |         | --- |
|     |        |             |         |         |     |
|     |        |             |         | 5       |     |
+  -1 +        +             +         +         +(1)  +
```

region (e.g., item 5: first referee's suitability). The final column (*S.1*) allows the applicant distribution to be viewed on the 5-point scale.

The MFRM operates under the assumption that multiple observations can be viewed as one theoretical construct (Bond & Fox, 2007). It appears that all of the criteria items used to assess prospective MEDC students (degree, writing ability, goals, work experience, referee quality, and suitability) fit within a unidimensional construct. *Unidimensional construct* is a measurement term used to describe situations when a single concept (e.g., suitability for MEDC program) underlies a set of items. A summary of the item characteristics and their facet statistics are located in Table 1. The Rasch model is often used to identify aspects of a particular facet that are not fitting; it does this by producing a set of Fit indices: Infit and Outfit values. According to Engelhard (1992), an acceptable range for Infit and Outfit statistics is 0.5 to 1.5. Typically, high Infit and Outfit statistics are indicators of multidimensionality within a facet. The work experience item had the highest Infit and Outfit value of 1.50, which is on the border of what would be considered an acceptable range. These statistics suggest that either the work experience item is functioning differently across the population of MEDC applicants or the raters are more erratic in the way they are viewing applicants' work experience. When the scale point difficulties were analyzed, our decision to analyze the items using the rating scale model rather than the partial credit model was supported.

Table 1
Items Measurement Report

Obsvd	Fair		Model	Infit		Outfit		
Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	Nu Items
3.8	2.96	.26	.08	0.6	–4	0.6	–4	1 Degree
3.8	3.02	.19	.08	1.0	0	1.0	0	2 Writing Ability
4.1	3.38	–.24	.09	0.9	0	0.9	0	3 Fit of Goals
3.4	2.40	.84	.07	1.5	4	1.5	4	4 Work Experience
4.4	3.82	–.87	.10	1.2	1	1.1	1	5 R1:Suitability
3.9	3.08	.13	.08	0.9	0	1.0	0	6 R1:Quality
4.3	3.66	–.63	.09	1.2	2	1.2	2	7 R2:Suitability
3.7	2.84	.38	.08	0.8	–2	0.8	–2	8 R2:Quality
4.2	3.46	–.35	.09	1.2	1	1.2	1	9 R3:Suitability
3.8	2.93	.30	.08	0.8	–2	0.8	–2	10 R3:Quality

Adj S.D. .48 Separation 5.67 Reliability .97
Fixed (all same) chi–square: 320.7 d.f.: 9 significance: .00
Random (normal) chi–square: 9.0 d.f.: 8 significance: .34

Items. The items "R1, R2, R3, Suitability" are based on the raters' judgements about the suitability of the referees that provided references. Most referees were rated as well suited to comment on the appropriateness of the applicants. Conversely, the raters interpreted the referees' comments relatively severely, producing Fair Average measures of 2.84 to 3.08. The Fair Average values represent the standardized average for an item that is produced after taking into account measures from the other facets (Linacre, 2012). Overall, the relevant degree, writing ability, and fit of personal goals with the nature of the counselling program were all average items situated closely around the 0 logit mark. The fixed (all same) chi-square of 320.7, *df* = 9 was found to be statistically significant ($p < .005$). This information solely suggests that the items differed in terms of difficulty, which indicates that the MEDC applicants were rated across a set of items designed for a broader range of ability levels. Furthermore, all of the item scores together produced a separation ratio of 5.67 and a reliability coefficient of 0.97, providing further evidence that each of the 10 items varied in difficulty.

Raters. The rater measurement report (Table 2) describes the behaviour of each of the 5 raters. Myford and Wolfe (2004) suggest that in situations that involve high-stakes decision-making, the Fit indices should be more stringent, adjusted to 0.8 to 1.2. We do not use this same stringency for items, as more variation is expected across items to appropriately capture the range of applicants' abilities. Ideally, when it comes to raters, we want no inconsistency at all, which is why the Fit indices are stricter. The Infit scores for the raters ranged from 0.80 to 1.30 and the Outfit scores ranged from 0.80 to 1.20. Although there is a wide range in the Fit values produced by raters, it would still be justified to state that the Infit and Outfit statistics fell within the generally accepted region. One student rater and one faculty rater both displayed ratings (Infit = 0.8; Outfit = 0.8) that would be "cramped" or "information poor," likely due to a central tendency effect. The other student rater demonstrated opposite rating behaviour (Infit score = 1.30 and Outfit score = 1.20), suggesting that the ratings given by this rater would be more erratic. Nonetheless, none of the raters' Infit or Outfit statistic values warranted removal of a rater or a re-marking of any applicant files.

Table 2
Rater Measurement Report

Obsvd	Fair		Model	Infit		Outfit		
Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	Nu Items
3.8	3.06	.15	.08	0.6	0	1.1	1	1 Faculty Counsellor (New)
3.9	3.18	.01	.08	0.6	–3	0.8	–2	2 Faculty Counsellor
3.9	3.11	.09	.09	0.6	0	1.0	0	3 Faculty Non-Counsellor
4.0	3.26	–.09	.07	0.6	–2	0.8	–3	4 Student Counsellor I
4.0	3.32	–.16	.08	0.6	4	1.2	3	5 Student Counsellor II

Adj S.D. .10 Separation 1.63 Reliability .73
Fixed (all same) chi–square: 18.1 d.f.: 4 significance: .00
Random (normal) chi–square: 4.0 d.f.: 3 significance: .26

The most severe rater (R1) had a measure of 0.15 and the most lenient rater (R5) had a measure of –0.16, producing a spread of ±0.16 logits, which is roughly one third of a logit difference between the most lenient and most severe rater. This disparity suggests that the raters were fairly homogeneous when it came to rating the applicants. However, the fixed (all same) chi-square of 18.1, *df* = 4, was statistically significant ($p < .005$), suggesting that there are notable rater differences. The separation ratio of 1.63 and the separation reliability coefficient of .73 also indicate that the raters were somewhat different in their view of the MEDC applicants across the 10 preadmission items. The Rasch interrater reliability (IRR) was .27 (1 - .73). Neither rater percent agreement nor a Cohen's kappa was calculated. For this study we used a fully crossed design, which means that all applicants were rated by both the most lenient and most severe raters. However, in situations where the design is not fully crossed, the reliability coefficient would need to be lower in order to ensure fairness for all the applicants. The lower the reliability coefficient, the more confident we can be in the results, given that a reliability coefficient of zero indicates no difference between any of the raters (Sudweeks, Reeve, & Bradshaw, 2005). Although the 2 student raters demonstrated more leniency and more variability according to fit statistics, these characteristics are of a magnitude that should not preclude these 2 student raters from further applicant rating events.

Generalizability Analysis

This G-Theory study is considered a two-facet (items by raters) fully crossed design. The applicants are not defined as a separate facet because they are the "object of measurement." The important information determined through a G-theory analysis is the variance components for the object of measurement, the facets, and the interactions between the facets and the object of measurement. These variance components are shown in Table 3.

Table 3
Estimated G Study Variance Components

	Components			
Source	Variance Component	%	*df*	SE
Persons (P)	0.0958	11.3	48	0.0224
Raters (R)	0.00427	0.5	4	0.00356
Items (I)	0.0665	7.9	9	0.0379
P * R	0.0775	9.2	192	0.00794
P * I	0.257	30.4	432	0.0217
R * I	0.0416	4.9	36	0.0110
P * R * I	0.303	35.8	1728	0.0103
G coefficient relative	0.86			
G coefficient absolute	0.85			

The variance component for the object of measurement (i.e., applicants) was 0.096, accounting for approximately 11% of the total variance. Ideally, this variance component should be higher, indicating the application process provided greater separation among the applicants. Typically, greater separation results in higher G coefficient values. The variance component for items reflects differences between each of the 10 items. The variance component for items was 0.067, accounting for approximately 8% of the total variance. Noteworthy for the consistency of decisions regarding applicants is the relatively high variance component (0.26) for the applicant-by-item, which revealed that the items ranked applicants differently. The fact that each applicant obtained relatively different scores on the different items makes selection decisions more difficult.

In contrast, the variance component for raters was 0.0043 (0.5%), indicating similar use of the rating scales by the raters. Similarly, the rater-by-item interaction was also a comparatively small percentage 4.9% (0.042) of the overall variance, indicating that the raters used the scale consistently on each item. Of importance for the consistency of the selection process of applicants, the variance component for the raters-by-applicants interaction was 0.078 (9.2%). This value suggests that the raters differed somewhat in the way they viewed and scored each applicant, reducing the consistency of the rating process of applicants.

Lastly, the residual variance component contains any other variability in the scoring process, including the interactions among applicants, raters, items, and any other facets not included. Ideally, the residual should be small compared to the other variance components. Unfortunately, the residual variance component accounted for the largest portion of variance (0.30 and 35.8% of the total variance). Almost 36% of the variability in applicants' scores cannot be fully accounted for, thus reducing the consistency in the application process.

The results from this study produced a G coefficient of 0.86, which indicates the amount of variance associated with each applicant's score based on the universal score. The Swiss Society for Research in Education Working Group (2010) suggests that an acceptable G coefficient is one that is greater than or equal to 0.80. According to these standards, this study produced a G coefficient that adequately supports the precision of the measures produced. The relative G coefficient was used as opposed to the absolute G coefficient because the behaviour of each individual applicant is viewed in relation to the behaviour of all the other applicants.

Discussion

The primary goal of our research was to assess the overall effectiveness of an applicant selection process for a MEDC program at one Canadian university, a selection process we believe is widely used in other Canadian institutions, but typically not subjected to scrutiny. This process necessarily included an analysis of rater behaviours, item characteristics, and to a lesser degree applicant behaviours. The MFRM and G-Theory were chosen because of their abilities to provide relevant and credible information about rater, item, and applicant consistencies.

Although comparisons between the Rasch and generalizability analyses were made, a detailed comparison was not provided, as direct comparisons illustrating that the effectiveness of these models has already been extensively researched (e.g., MacMillan, 2000).

The Items Used for Selecting Applicants

Based on the MFRM, it appears the items differed in terms of difficulty. The Fit indices of the items also suggest the presence of an underlying unidimensional construct. The most difficult item was the work experience item (0.84 logits); this was not entirely unexpected given the range of applicants' relevant experience. A vast majority of the MEDC applicants had recently completed an undergraduate degree and did not yet have the opportunity to gain relevant work experience in a helping profession. The least difficult item was finding one suitable referee to support the applicant pursuing admission into the MEDC program (-0.87 logits).

Furthermore, the MFRM also identified potentially problematic items. Items are problematic when they function differently across a population or when their level of difficulty is not appropriately matched to a person's ability. For example, the work experience item was potentially misfitting based on the Mean Square Fit values (Infit = 1.5 and Outfit = 1.5). Subsequently, we linked these values to the full-time and part-time status of applicants and discovered that the full-time and part-time applicants were separated on this measure. The work experience item was functioning differently across the pool of MEDC applicants and was biased toward the full-time applicants who typically did not have as much work experience as the part-time applicants. In contrast, the relevant degree item (0.26 logits) had extremely low Infit (0.6) and Outfit (0.6) statistics, suggesting this item was too predictable. In other words, this item was scored with a level of consistency that was far superior to the other items. Hence, a single rater—for example, an administrative assistant or one of the current raters who was previously taught how to use the scale for this item—could likely score this item. The high separation ratio (5.67), which indicates the differences among the items were more than five times greater than the errors associated with the measurement model, and reliability index (0.97) for the items suggests the items measured different aspects of the applicants' overall admission and captured different traits.

The G-Theory analysis provided further evidence regarding the application process for this particular MEDC program. Specifically, the G-Theory analysis demonstrated that the items varied slightly by showing that 7.9% of the variance was due to the items facet. Of some concern is the relatively high applicant-by-item interaction (30.4%). There is a relatively large amount of variability that can be attributed to differences in each applicant's scores across the different items. The fact that each applicant received diverse scores on different items suggests that some items are functioning differently for the various groups (i.e., full-time and part-time) of MEDC applicants.

The combined MFRM and G-Theory analyses indicate that the items used in the application process do not provide a fully consistent measure of each applicant. One likely explanation is that the items appear to measure somewhat different traits among the applicants. Hence, our findings highlight the importance of using different items to select the most desirable applicants. Using this approach (i.e., conjunctive, which has a unique standard defined for each item) for admissions purposes requires applicants to demonstrate excellence across all the preadmissions criteria. However, in this particular case, as well in most admissions processes, a compensatory model was used. A compensatory model combines the results from each of the items to produce a single score for each applicant and that combined score is subsequently used with respect to making admission decisions. The issue that arises when items are combined is that some of the strengths and weaknesses of applicants are hidden as the candidate is now assessed holistically. By combining preadmissions items, applicants' high score on one item can compensate for their poor score on another item, which is problematic as it does not necessarily lead to the best possible candidates being accepted.

The Raters Who Reviewed the Applicants

Five raters participated in this study: 3 faculty members and 2 students. As far as the rater analysis was concerned, the Rasch was most informative in providing information about how each rater behaved individually. The rater measurement report indicated a 0.31 logit spread among the 5 raters. A 0.31 logit spread is relatively low, considering the diversity of knowledge and experience among the raters. The separation ratio (1.63) and reliability index (.73) for raters suggest differences, albeit relatively small ones, among the 5 raters. The Rasch analysis showed that the most severe rater (0.15 logits) was a new faculty member who had a counselling background but no previous experience with the admissions process at this particular institution. The next most severe rater (0.09 logits) was another faculty member, one who did not have a counselling background but had considerable experience with this process. The faculty member who had a counselling background and was familiar with the application process was situated in the middle of the 5 raters (0.01 logits). Overall, the 2 student raters were the most lenient of all the raters (-0.09 and -0.16 logits), with the first being overly constrained and the other being somewhat erratic according to the fit statistics.

The G-Theory analysis attributed 0.5% of the variance components to raters, which indicated that the raters scored the applicants fairly consistently as there was little variation due to rater differences. The small amount of variance accounted for by the rater-by-item interaction (4.9%) suggested that all of the raters used the rating scale similarly for each item. These results support the use of a less than fully crossed rater-by-applicant design. However, the 9.2% variance component for the rater-by-applicant interaction may indicate otherwise. This rater-by-applicant G-Theory variance component shows that the raters slightly differed in how they scored various applicants.

Conclusions

After an exploration of the relationship between the MFRM and G-Theory, it appears that each methodology has its prevailing strengths and weaknesses. The strengths of the Rasch model included greater detail when focusing on the individual characteristics of each facet and supplying error indicators for each element, as well as a remarkable ability to handle small sample sizes. Some of the weaknesses of the Rasch model relate to its simplicity. The Rasch model is not overly complicated, which has some researchers convinced that it is not a viable model. Also, the lack of concrete rules relating to sample size and fit statistics has been a source of frustration. The strengths of the G-Theory included the ability to provide variance components for each facet's main effect and all possible interactions, the freedom to make relative or absolute decisions, and the decision studies feature (not included in this study), which displays reliability measures for various future designs. G-Theory is also able to produce group statistics for various facets, which can be generalized to the broader population. Some of the weaknesses of G-Theory have to do with its inability to provide specific details about individual characteristics within a facet.

According to the Rasch and G-Theory analyses, the items used to evaluate MEDC applicants were fitting in their ability to measure various aspects of a unidimensional construct. The range in item difficulties (0.84 to -0.87 logits) suggests a sufficient spread in the items measuring counselling applicants. However, the level of item difficulty in relation to applicants' ability suggest that some applicants were outperforming the admission items, as the items appeared to be targeting more applicants on the lower half of the population (see Figure 1). As indicated by the G-Theory results, the amount of variance attributed to the main effect of applicants was relatively low (11.3%), suggesting some homogeneity in the sample. The reasoning for this correspondence is that in homogeneous populations, raters have a harder time differentiating between applicants and thus usually produce lower G coefficient values. This is a common issue with admission criteria in very competitive graduate and professional programs, as most applicants are highly qualified and suitable based on the admissions criteria. Where this becomes challenging is deciding whether or not to alter the difficulty levels of items to hopefully better separate the applicants. In some cases (e.g., relevant degree), this might require all applicants seeking admission into the MEDC program to have at least a master's, if not a PhD prior to being accepted. Not only are these standards unrealistic, but they do not actually guarantee that those best suited for the counselling profession will be admitted. The other alternative to dealing with the discrepancy between applicants' ability levels and item difficulties would be to revise the rating scale to reflect the high quality of applicants such programs receive. In considering the rating behaviour of the participants on the selection committee, both the Rasch and G-Theory analyses suggested that the raters were consistent as individuals and suitable as a group (0.31 logit spread, .73 reliability index, and $p < .005$; 0.5% variance for rater main effect, 9.2% variance for applicant-by-rater interaction, and 4.9% for rater-by-item interaction). Besides some notable differences between the faculty and student raters, it appears that the raters used in the MEDC application selection process showed minimal variation and were fairly consistent in their behaviour.

In conclusion, the MFRM and G-Theory both have prevailing strengths. Each methodology was designed with an idea of the optimal conditions that would warrant the use of that particular methodology. Research in the area of measurement requires researchers to make judgements as to whether the measurement context is appropriately suited to the methodology. Sometimes one methodology is not sufficient to adequately measure all of the questions that a researcher has. Therefore, with any analysis, it may be necessary to find two or more measurement models that can be combined to make the most out of the information situated within the data.

Limitations of the Design

The G-Theory analysis produced a large residual variance component of 35.8% for the 49 applicants, which is a source of concern considering the design and the amount of information that was accumulated by simultaneously employing two different measurement models. The small sample size makes it unclear as to what unexplained factors are responsible for such a high variance component. Furthermore, no follow-up interviews were conducted with any of the raters that may help account for such differences. That this was the first time any assessment of the MEDC admissions criteria occurred also makes it difficult to support specific claims about the reliability of the application selection process.

Recommendations for Future Research

The application selection committee could investigate and experiment with other potential items, such as adding a supplementary item to settle the variability of the work experience item or removing the relevant degree item from the rating scale and having the relevant degree coded by one person. Also, a reliability study including an examination of rater drifts (i.e., the rating behaviour and patterns that change over time) would be worthwhile. One of the most common facets analyzed with both Rasch analysis and G-Theory is occasions (i.e., ratings over time), which was not utilized in this particular study. As the data for this study were gathered on only one occasion, the opportunity to examine item difficulty, rater behaviour, or applicant quality over any period of time was not possible. Finally, a validity study tracking the successful MEDC applicants throughout their program would also provide valuable information about how well the application selection process is currently operating.

Enhancing the Assessment of MEDC Applicants

Every year admissions processes have the difficult task of identifying the best applicants for a particular MEDC program. One of the aspects that makes the application selection process challenging is the unique rating behaviour of individuals, which is often a main source of variability within assessments. Undesired variability in the assessment of MEDC applicants is problematic as it can threaten the validity and fairness of an admissions process. Therefore, regularly collecting admissions data and monitoring the assessment of MEDC applicants for both applicant quality and rater consistency is beneficial in enhancing the MEDC admission process used at any institution.

This article described two different approaches that individuals can use in order to evaluate the selection process of MEDC applicants at their own institutions. Through the use of G-Theory, individuals can obtain information about how much variance is attributable to raters, items, and the applicants themselves. Alternatively, the MFRM showed how to obtain specific information about individual raters, items, or applicants. Furthermore, the MFRM can correct for differences in severity among raters. Being able to identify rater differences or potentially misfitting items is most advantageous, as the information obtained about raters or items can be used to make adjustments and facilitate communication among individuals at a particular institution.

For example, during the assessment of MEDC applicants at the participating institution, a conversation occurred among the raters regarding the values inherent in the MEDC program (e.g., preparing practitioners to work in underserviced regions). This conversation was facilitated by the results of the Rasch analysis, which highlighted noteworthy differences observed between faculty and student raters on the work experience admissions item. Rater agreement in the selection of MEDC applicants is not always necessary, as different raters might distinguish between applicants using various approaches; however, it is imperative that raters are given an opportunity to communicate how they are uniquely seeing individual MEDC applicants.

Graduate and professional programs with sound admissions processes are generally the ones that have a well-articulated purpose for their program and provide individuals involved in applicant selection the opportunity to clearly communicate what attributes, characteristics, and experiences are valued in prospective students applying to a particular MEDC program. Assessing the admissions processes of individual MEDC programs helps not only to ensure that admissions processes are fair and accurate, but also promotes the selection of applicants who are most suitable for a particular program and hopefully one day the counselling profession.

Acknowledgements

The authors wish to thank Dr. Serge Demers and Dr. Yvon Cloutier for their assistance and editing of French translations.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. doi:10.1007/BF02293814

Atilgan, H. (2008). Using generalizability theory to assess the score reliability of the special ability selection examinations for music education programmes in higher education. *International Journal of Research & Method in Education, 31*(1), 63–76. doi:10.1080/174372708011919925

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed). Mahwah, NJ: Lawrence Erlbaum.

Canadian Counselling and Psychotherapy Association. (2013). *Graduate programs*. Retrieved from http://www.ccpa-accp.ca/en/aucc_program/

Chang, W., & Chan, C. (1995). Rasch analysis for outcome measures: Some methodological considerations. *Archives of Physical Medicine and Rehabilitation, 76*(1), 934–939. doi:10.1016/S0003-9993(95)80070-0

Corey, G., Corey, M. S., & Callanan, P. (2007). *Issues and ethics in the helping professions* (7th ed). Belmont, CA: Thomson/Cole.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221. doi:10.1207/s15434311laq0203_2

Engelhard, G. (1992). The measurement of writing ability with a many-facet Rasch model. *Applied Measurement in Education, 5*(3), 171–191. doi:10.1207/s15324818ame0503_1

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-facet Rasch model. *Journal of Educational Measurement, 31*(2), 93–112. doi:10.1111/j.1745-3984.1994.tb00436.x

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43–58. doi:10.1111/j.1745-3984.2009.01068.x

Homrich, A. M. (2009). Gatekeeping for personal and professional competence in graduate counseling programs. *Counseling and Human Development, 41*(7), 1–23.

Kieffer, K. M. (1999). Why generalizability theory is essential and classical test theory is often inadequate. *Advances in Social Science Methodology, 5*(1), 149–170.

Kim, S. C., & Wilson, M. (2010). A comparative analysis of the rating in performance assessment using generalizability theory and the many-facet Rasch model. In M. L. Garner, G. Engelhard, W. P. Fisher, & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 1, pp. 304–327). Maple Grove, MN: JAM Press.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA.

Linacre, J. M. (1996). *FACETS: A computer program for analysis of examinations with multiple facets, version 3.03*. Chicago, IL: MESA.

Linacre, J. M. (2010). *Rasch measurement: Core topics*. Retrieved from http://courses.statistics.com/index.php3

Linacre, J. M. (2012). *Facets computer program for many-facet Rasch measurement, version 3.70.0*. Beaverton, OR: Winsteps

Linacre, J. M., Wright, B. D., & Lunz, M. E. (1990). *A facets model for judgmental scoring*. Retrieved from http://www.Rasch.org/memo61.htm

MacMillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *Journal of Experimental Education, 68*(2), 167–190. doi:10.1080/00220970009598501

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. doi:10.1007/BF02296272

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 460–517). Maple Grove, MN: JAM Press.

Nelson, K. W., Canada, R. M., & Lancaster, L. B. (2003). An investigation of nonacademic admission criteria for doctoral-level counselor education and similar professional programs. *Journal of Humanistic Counseling, Education, and Development, 42*(1), 2–13. doi:10.1002/j.2164-490X.2003.tb00164.x

O'Neill, T. R. (1999). Adjusting for rater severity over time. *Popular Measurement, 47*(1), 46–47.

Oosterveld, P., & ten Cate, O. (2004). Generalizability of a study sample assessment procedure for entrance selection for medical school. *Medical Teacher, 26*(7), 635–639. doi:10.1080/01421590400004874

Pass, L. E., & Scherer, S. E. (1979). Toward more adequate selection criteria: A case study of graduate counselling admissions. *Canadian Journal of Counselling and Psychotherapy, 13*(3), 127–130. Retrieved from http://cjc-rcc.ucalgary.ca/cjc/index.php/rcc/article/view/1942

Pedersen, G., Hagtvet, K. A., & Karterud, S. (2007). Generalizability studies of the global assessment of functioning: Split version. *Comprehensive Psychiatry, 48*(1), 88–94. doi:10.1016/j.comppsych.2006.03.008

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

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

Smith, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement, 64*(4), 617–639. doi:10.1177/0013164404263876

Stone, G. E., Beltyukova, S., & Fox, C. M. (2008). Objective standard setting for judge-mediated examinations. *International Journal of Testing, 8*(2), 180–196. doi:10.1080/15305050802007083

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(1), 239–261. doi:10.1016/j.asw.2004.11.001

Swiss Society for Research in Education Working Group. (2010). *EDUG user guide, version 6.0*. Neuchatel, Switzerland: Edumetrics.

Winne, P. H., Nesbit, J. C., Kumar, V., & Hadwin, A. F., Lajoie, S. P., Azevedo, R. A., & Perry, N. E. (2006). Supporting self regulated learning with g-study software: The learning kit project. *Technology, Instruction, Cognition and Learning, 3*(1), 105–113.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35–51.

About the Authors

Stefanie Sebok is a third-year PhD student specializing in measurement, assessment, and evaluation at Queen's University. Her main interests include exploring the rating behaviour of assessors, particularly in high-stakes assessment situations. She also works as a personal counsellor at Queen's University.

Peter MacMillan is an associate professor and Chair of the School of Education at the University of Northern British Columbia. His research interests include applications of Rasch Modelling for educational and health data analyses.

Address correspondence to Stefanie Sebok, Queen's University, LaSalle Building Room 209, 146 Stuart Street, Kingston, ON, Canada, K7L 3N6; e-mail <stefanie.sebok@queensu.ca>