Adv in Health Sci Educ (2014) 19:71–84
DOI 10.1007/s10459-013-9463-7

Psychometric properties of the multiple mini-interview used for medical admissions: findings from generalizability and Rasch analyses

Stefanie S. Sebok · King Luu · Don A. Klinger

Received: 14 December 2012 / Accepted: 13 May 2013 / Published online: 25 May 2013

Abstract The multiple mini-interview (MMI) has become an increasingly popular admissions method for selecting prospective students into professional programs (e.g., medical school). The MMI uses a series of short, labour intensive simulation stations and scenario interviews to more effectively assess applicants' non-cognitive qualities such as empathy, critical thinking, integrity, and communication. MMI data from 455 medical school applicants were analyzed using: (1) Generalizability Theory to estimate the generalizability of the MMI and identify sources of error; and (2) the Many-Facet Rasch Model, to identify misfitting examinees, items and raters. Consistent with previous research, our results support the reliability of MMI process. However, it appears that the non-cognitive qualities are not being measured as unique constructs across stations.

Keywords Multiple mini-interview · Medical admissions · Rasch measurement · G theory

Introduction

Medical school students are typically assessed through objective structured clinical examinations (OSCE; e.g., Harden and Gleeson 1979) as part of their medical education and subsequent licensure examinations. The OSCE procedure requires students to demonstrate their clinical skills across a series of simulated medical scenarios. In 2001, the medical school at McMaster University in Canada used the OSCE model to pilot a new

S. S. Sebok (✉) · K. Luu · D. A. Klinger
Faculty of Education, Queen's University, Kingston, Ontario, Canada
e-mail: stefanie.sebok@queensu.ca

K. Luu
e-mail: king.luu@queensu.ca

D. A. Klinger
e-mail: don.klinger@queensu.ca

type of admissions interview—the multiple mini-interview (MMI; Eva et al. 2004), which was subsequently incorporated as part of the University's medical admissions assessment. Since then, the MMI has become an increasingly common admissions tool at several other Canadian and international medical schools. Further, other highly competitive professional health related programs including dentistry, pharmacy, and nursing have employed the MMI for admissions screening. Rather than the traditional panel interview used to select applicants for admission to medical school, the MMI requires applicants to participate in a series of mini-interviews that address specific dilemmas or situations. For example, applicants may interact with a standardized patient, or be engaged in a discussion about a particular ethical issue. Experienced medical professionals, and in some cases upper-year medical students, assess applicants' performance at each station. Applicants rotate through a number of stations, completing between seven and ten tasks.

The MMI was first introduced in an effort to measure important, non-cognitive traits such as communication and critical thinking that traditional interviews did not seem to assess. Further, proponents of the MMI claim it reduces interviewer and situational biases that exist during individual panel interviews (e.g., Eva et al. 2004). Eva and his colleagues have conducted a number of studies on the MMI program at McMaster University. Among their findings, the MMI: (1) had comparable validity evidence to MCAT scores and applicant GPA; (2) was better at predicting subsequent OSCE performance and grade point average (GPA) than traditional assessment tools; and (3) successfully determined clinical clerkship performance and national licensing examination performance (Eva et al. 2004; Reiter et al. 2007). Currently, McMaster University places a greater emphasis on MMI results than students' MCAT results.

Although the MMI has been increasingly used in numerous medical schools to supplement their individual admission processes, the body of research exploring the MMI remains relatively thin outside of the work done at McMaster University. The MMI appears to have high reliability, and differentiates qualities amongst applicants (Dodson et al. 2009); however, there has been little research about the consistency of applicant performance across the stations, or of raters' abilities to differentiate amongst the non-cognitive traits. As an example, Roberts et al. (2008) found that the largest source of error was related to rater subjectivity and that low correlations existed in applicant performance amongst stations. These findings suggest that there exist important rater (interviewer) and context effects that need to be further explored.

The purpose of the current study was to examine the psychometric properties of the MMI as employed at another Canadian medical school. The application process for this medical program is highly competitive with approximately 100 applicants being granted admission each year. This research study used MMI data from 455 medical school applicants who were short-listed for admission. Along with descriptive analyses, Generalizability theory (G theory) and the Many-Facet Rasch Model (MFRM) were used to estimate the generalizability of the MMI, various sources of variance, and presence of psychometrically misfitting applicants, items, and raters.

Generalizability theory

G theory is used to assess the consistency or dependability of scores over randomly parallel replications of a particular measurement. It can be used to estimate the magnitude of various sources of error in observed scores as well as the relationships among such sources. G theory also provides a mechanism for optimizing the generalizability (reliability) of other measurements (e.g., Brennan 2001; Shavelson and Webb 1991). Each set of scores is a sample from a universe of admissible observations, which consists of all possible observations on an object of measurement (i.e., the factor we are interested in, usually persons). Each characteristic of measurement is defined as a facet (Webb et al. 2006). Variability in the facets can be potential sources of error, lowering the generalizability of the results. The primary advantages of G theory include the ability to estimate the relative importance of multiple sources of error in a single analysis for a given situation. It can also provide information useful in optimizing the design of a measurement procedure. The generalizability coefficient ($\boldsymbol{E}(\rho^2)$) provides an estimate of the generalizability of scores for norm-referenced interpretations; for example, the selection of applicants for admission to medical school. $\boldsymbol{E}(\rho^2)$ encompasses all sources of error that influence the relative standing of persons (relative error variance).

A G study estimates the variation due to the object of measurement and its associated facets, while a decision (D) study uses the information from the G study to find the optimal conditions for a particular measurement design (Shavelson and Webb 1991). An optimized design minimizes the undesirable sources of error and maximizes generalizability. To ensure the best results, a universe of generalization needs to comprise the whole collection of possible observations over the population we wish to generalize.

Many-Facet Rasch Model

The Rasch model employs the principles of interval measurement to objectively measure data. This involves taking raw scores of ordinal nature and performing a series of logarithmic transformations to produce data that support linearity (Linacre 2010). Georg Rasch (1901–1980) created the first Rasch model that some researchers in the field of measurement still use today. The traditional Rasch model was designed for dichotomously scored data and focused on two aspects: person ability and item difficulty. The MFRM was developed by Linacre (1989) and expanded upon the original Rasch model to include multiple facets (e.g., raters or occasions).

Fox and Jones (1998) explain that a Rasch analysis produces a set of Infit and Outfit values for each person as well as for each item. These fit statistics are useful to researchers because they provide information that identifies unexpected responses. Fit statistics have a mean-square value of 1.0 and can range anywhere from zero to infinity (Linacre 1995). Mean-square values that are greater than 1.0 are labeled “underfit,” while mean-square values less than 1.0 are labeled “overfit” (Linacre 2010). Outfit mean-square statistics provide information about unexpected response patterns. It would be more sensitive to situations where a person with a high ability answered a really easy item incorrectly or a person with a low ability answered a very difficult item correctly. Infit mean-square statistics provide information about the unexpected response patterns by examining the region where a person’s ability generally is the same as the item difficulty (Wright and Linacre 1994).

Method

Sample

The data for the current study were obtained from the 2011 medical admissions process—the first year the MMI was used to support admission decisions at this particular medical school. Due to security concerns, the individual stations cannot be described in more detail;

however, they can be broadly described as a series of scenario based mini-interviews (similar to the process used at McMaster University), combined with simulations more similar to the traditional OSCE. During each simulation or mini-interview, the applicants demonstrate their ability to handle a moral, ethical, or situational issue. Two raters scored the applicant at each station: a faculty member and an upper-year medical student. Further differences include the use of seven mini-interview stations, with an eighth station consisting of a traditional interview. Further, different non-cognitive qualities (*attributes*) are measured through this MMI compared to the one employed at McMaster.

All applicants were rated using a 9-point scale for each of the attributes within a station. The 9-point scale was not described in detail by the test developers; rather, it was anchored at three points (1 = 'Poor', 5 = 'Satisfactory', 9 = 'Outstanding'). For each attribute, the applicant received two scores: one score from a faculty rater and another from a student rater.

Procedure

GENOVA software was used to calculate all G coefficients (Crick and Brennan 1983). Applicants with any missing scores were removed prior to analysis so that the design remained balanced, resulting in a sample of 444 applicants. In this study, a nested design was used since each of the eight stations was rated by a different group of two raters. Given that each station was comprised of an unequal number of attribute scores, we modified the scores used for the G study. A station score was calculated for each applicant by averaging the attribute scores for each rater within a station, resulting in a faculty score and student score for each station. As admission decisions from the MMI are based on rank order, we focused on the determination of $\boldsymbol{E}(\rho^2)$ (see Eq. 1) and the relative error variance components.

$$\boldsymbol{E}(\rho^2) = \sigma^2(p)/\sigma^2(p) + \sigma^2(\delta) \tag{1}$$

The model underlying the scores for the MMI in our study is a two-facet $p \times (r{:}s)$ random-effects nested design. The random-effects design assumes that the stations and raters are randomly selected from a universe of possible stations and raters. The variance components in the G study for this design are:

- $\sigma^2(p)$ applicants
- $\sigma^2(s)$ stations
- $\sigma^2(r{:}s)$ raters within stations
- $\sigma^2(ps)$ applicants by stations
- $\sigma^2(pr{:}s)$ applicants by raters within stations

The variance components for D studies are the same, except that the notation is capitalized when the facet is fixed in the universe. For example, $\sigma^2(r{:}s)$ becomes $\sigma^2(R{:}S)$. Relative error variance ($\sigma^2(\delta)$) is associated with the use of an individual's observed deviation score as an estimate of the individual's universe deviation score. It is the sum of all variance components that include the object of measurement (universe score) and another facet in the D study. For the $p \times (R{:}S)$ random-effects nested design,

$$\sigma^2(\delta) = \sigma^2(pS) + \sigma^2(pR : S) \tag{2}$$

In the case of the D study, the variance components are estimates of error variance, where the number of raters and the number of stations are facets of interest and can be altered to investigate how changes to the facets can improve score reliability. The following D study designs were conducted: univariate, $p \times (R{:}S)$, with both raters and stations random and univariate, $p \times (R{:}S)$, with raters random and stations fixed. A variety of raters and station combinations were done for the former to find the minimal design that would result in a $\boldsymbol{E}(\rho^2)$ of at least 0.80.

The Many-Facet Rasch analysis was conducted on the complete set of 455 applicants using FACETS software, version 3.68.1 (Linacre 2011). Applicants with 0 scores were kept in the analyses because Rasch models are able to determine an expected score value. The MFRM is an extension of the original Rasch measurement model as it goes beyond person ability and item difficulty to measure other factors that interact with a testing situation. The formula for the MFRM can be described as follows:

$$\ln\left(\mathrm{P}_{nijk}/\mathrm{P}_{nij(k-1)}\right) = \mathrm{B}_n - \mathrm{D}_i - \mathrm{C}_j - \mathrm{F}_k$$

where P_{nijk} = probability of examinee n being graded k by judge j on item i, $\mathrm{P}_{nij(k-1)}$ = probability of examinee n being graded $k - 1$ by judge j on item i, B_n = performance measure of examinee n, D_i = difficulty of item i, C_j = severity of judge j, F_k = difficulty of grading Step [category] k relative to Step [category] $k - 1$, (Lunz et al. 1990).

Initial analysis

The eight non-cognitive attributes intended to be measured by this particular MMI were communication, critical thinking, effectiveness, empathy, integrity, maturity, professionalism, and resolution, which were described as items within each station. At each station of the MMI, applicants were scored on one to four attributes (see Table 1). Table 1 provides the descriptive results for the attribute scores at the eight MMI stations. Eleven applicants were removed because their information was incomplete (e.g., not scored by one of the raters), resulting in a sample size of 444 applicants. The results are separated by faculty and student raters and correlations are provided comparing the associations between the two raters at each station. As presented in Table 1, the faculty and student raters' scores were relatively similar across the stations, with student raters tending to give slightly higher scores; however, this was not consistent across every station. The variability in scores was similar for each of the station scores. The faculty and student rater attribute scores were moderately correlated, ranging from 0.41 to 0.69, which are relatively common for such rating scales (e.g., Thorndike 2005). The lowest average scores were given for the 'professionalism' item in Station 4, indicating that it was the most difficult for applicants. Station 8, which was most similar to the traditional interview, was the easiest station.

Generalizability analyses and results

The variance components for the $p \times (r{:}s)$ random effects nested design are presented in Table 2. The estimated variance component for applicants was 0.47, which accounted for 16.3 % of the total variance. The variance component for applicants is important as it represents the differences between applicants in how they scored on the overall MMI, it also illustrates the extent to which applicants varied in ability. The largest variance component was *ps,* the interaction between applicants and stations (1.21, 41.7 %). This indicates that there were differences in the scores obtained by applicants at each station. The ps interaction reduces the consistency of scores as it shows the relative standing of

Table 1 Descriptive information on station scores from faculty and student raters

	Faculty raters		Student raters		*r*
	M	*SD*	*M*	*SD*	
Station 1					
Communication	5.91	1.6	5.76	1.9	0.48
Critical thinking	5.77	1.7	5.61	1.8	0.41
Maturity	5.89	1.7	5.78	1.8	0.42
Station 2					
Communication	5.79	1.8	5.86	1.8	0.69
Critical thinking	5.40	2.0	5.57	1.9	0.65
Station 3					
Communication	5.57	1.9	5.53	2.0	0.48
Critical thinking	5.45	1.9	5.52	1.9	0.55
Maturity	5.46	1.9	5.69	1.8	0.51
Station 4					
Effectiveness	5.60	1.8	5.63	1.6	0.52
Empathy	5.62	1.8	5.74	1.8	0.48
Professionalism	4.52	2.5	5.01	2.3	0.58
Resolution	5.60	1.9	5.85	1.9	0.57
Station 5					
Integrity	5.50	1.6	5.86	1.7	0.51
Maturity	5.42	1.7	5.69	1.9	0.54
Station 6					
Communication	5.74	1.9	5.75	1.9	0.57
Empathy	5.52	1.9	5.56	2.1	0.60
Critical thinking	5.80	1.7	5.83	1.7	0.59
Station 7					
Critical thinking	5.50	1.8	5.49	1.9	0.58
Professionalism	5.83	1.6	5.71	1.6	0.41
Maturity	5.79	1.7	5.77	1.8	0.45
Station 8					
Communication	6.12	1.9	6.21	1.8	0.56

applicants differed across stations. There was also a large residual error variance (*pr:s*) of 1.17, or 40.3 % of the total variance, suggesting that a large proportion of the effects are confounded by the interactions between applicants, raters, stations, and other unexplained sources of error. In contrast, there was little variability amongst stations (0.04, 1.3 %) and raters nested within stations (0.01, 0.4 %), indicating the stations had similar variability and the raters were scoring examinees consistently.

The results from the G study were subsequently used to conduct a series of D studies. D studies allow us to estimate the generalizability of the current the MMI admissions procedure as well as the generalizability that could be expected to obtain through the use of more stations or raters. Based on these variance components and the current random-effects design of eight stations with two raters per station, the G coefficient $\boldsymbol{E}(\rho^2) = 0.68$. If the number of stations was increased to nine (two additional observations), or the

Table 2 G Study variance components of the $p \times (r{:}s)$ random effects model for the 2011 MMI

Effect	df	MS	Variance component (σ^2)	% of overall variance
p	443	11.18	0.47	16.3
s	7	40.73	0.04	1.3
r:s	8	5.85	0.01	0.4
ps	3,101	3.60	1.21	41.7
pr:s	3,544	1.17	1.17	40.3

Table 3 G and D Study variance components, error variance and generalizability coefficient estimates for $p \times (R{:}S)$ random-effects design

n'_S n'_R	7 2	7 3	7 4	8 2	8 3	9 1	9 2	9 3	15 2
p	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47
S	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
R:S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
pS	0.17	0.17	0.17	0.15	0.15	0.13	0.13	0.13	0.08
pR:S	0.08	0.06	0.04	0.07	0.05	0.13	0.07	0.04	0.04
$\boldsymbol{E}(\rho^2)$	0.65	0.67	0.69	0.68	0.70	0.64	0.70	0.73	0.81

n'_S number of stations, n'_R number of raters

number of raters is increased to three per station (eight additional observations), $E(\rho^2)$ increases to 0.70. Since the value for $\sigma^2(ps)$ is relatively large, increasing the number of stations would lead to a greater reduction in error variance compared to adding more raters per station (see Table 3). This increase would also be less costly as an increase of one station only requires two more raters, while each increase in the number of raters per station requires eight more raters.

If a G coefficient of 0.80 or greater is desired (i.e., for relative decisions), the design would need to have a minimum 15 stations (with two raters at each station). A similar value could also be obtained with 14 stations and 3 raters per station (results not shown), but as mentioned previously, this would be a more costly alternative requiring 12 additional raters. It is not possible to obtain a value for $\boldsymbol{E}(\rho^2)$ of 0.80 with only eight stations, regardless of the number of raters used.

Alternatively, the stations used in the MMI could be fixed at eight. This requires the assumption that the current MMI stations used are representative of the entire set of possible stations that exist. Under this assumption, the G coefficient $\boldsymbol{E}(\rho^2) = 0.90$ with two raters per station. The coefficient is quite high because the universe of generalization is small, thereby limiting generalizability. Not surprisingly, increasing the number of raters gradually brings the coefficient closer to one. In this fixed design, the use of a single rater per station reduces $\boldsymbol{E}\rho^2$ to 0.81.

Many-Facet Rasch analyses and results

The Wright Map (Fig. 1) provides an overall picture of the applicants, raters, and items measured at each station. Each facet—medical applicants, raters, and items, in terms of

Measure	Applicants	Raters	Items	Scale
2				**(9)**
				8
	.			
	.			
1	.			
	*			
	**			
	*.			**7**
	***.			
	***.			
	*****			
	*******.		11	**6**
	**********			
	*********		5 7 18	
0	*******.	Faculty Student	2 6 8 9 10 12 13 14 15 16	
			19 20	
	*****.		1 3 4 17	**5**
	****.		21	
	*.			
	*.			**4**
	.			
	.			
				3
-1				**(1)**

Fig. 1 Wright variable map displaying relationships for medical school applicants. Each *asterisk* represents 7 medical school applicants

ability, severity, and difficulty—is placed on the same measurement scale (i.e., the logit scale). The first column (Measure) represents the logit scale used to measure the various facets in this study. A logit is a unit of measurement that results from a log-odds transformation. Characteristically, 0 is allocated as the mean. Positive values represent higher ability, more severity, or greater difficulty, while negative values represent lower ability, more leniency, or lesser difficulty. The second column shows the distribution of medical applicants. Most of the applicants are situated in the 0–1 region on the logit scale, indicating that they were all fairly proficient. The third column contains the rater facet. Both the faculty and student raters are positioned around the 0 logit mark, which affirms that overall the faculty and student raters were equivalent in terms of their severity. The fourth column represents the items and their associated attributes (for detailed list of these items see "Appendix"). The more difficult items are located in the positive logit region, while the less difficult items are located in the negative region. Hence, in alignment with our initial analysis, the professionalism item at Station 4 (Item 11) was the most difficult, while the communication item at Station 8 (Item 21) was the easiest. The final column (Scale) shows the ratings on the 9-point scale and denotes the medical applicants, raters, and items along a universal continuum. Almost all of the applicants had scores between 4 and 7, which was aligned with the placement of the majority of the applicant and all of the items. This information, coupled with the absence of category 2, suggests that the 9-point scale was not being fully used.

Based on the logit measure for the raters it appears that there is a 0.04 logit difference between the faculty and student raters, with the faculty raters being slightly more severe. All of the Infit and Outfit statistics for the raters fell within an acceptable range. These results are consistent with the generalizability analyses that suggested that the raters were fairly homogeneous; however, the reliability coefficient of 0.84 suggests that there were some important differences in severity between the faculty and student raters. In this particular case, a smaller reliability coefficient is desired, as lower values indicate less difference between the raters; when a reliability coefficient is zero, the raters are for all intents and purposes interchangeable.

The information presented in Table 4 represents the rating behaviour of faculty and student raters for the 455 medical school applicants on eight stations, twenty-one items, and eight constructs. Reliability coefficients for medical school applicants, raters, and items were also separately calculated for each of the eight stations. The applicant reliability coefficients for the medical school applicants across the eight stations range from 0.84 to 0.92, which indicates heterogeneity among the applicants. Synonymous to G theory, where a large amount of variance for persons is desired, higher reliability coefficient values are preferred for applicants because larger values indicate the raters were able to separate the applicants across items. In contrast, the rater reliabilities should be low across the stations, as this would indicate that faculty and student raters were similar at each station. For example, Stations 4 and 5 had rater reliability values of 0.94 and 0.96, which suggest the faculty and student raters at these particular stations were extremely heterogeneous in their ratings of the medical school applicants. On the contrary, Station 6 produced a reliability coefficient of 0.00, suggesting that there was no difference between the raters at this station.

Lastly, the item reliabilities should also be high as the items represent different non-cognitive attributes. For instance, at station 6 there were three attributes being measured: communication, critical thinking, and empathy. The reliability coefficient of 0.91 indicates that those three attributes are distinct; furthermore, the faculty and student raters assessed

Table 4 Reliability coefficients report by station

Station	Applicant reliability values	Rater reliability values	Item reliability values
s_1	0.90	0.80	0.76
s_2	0.92	0.60	0.97
s_3	0.91	0.64	0.07
s_4	0.89	0.94	0.99
s_5	0.89	0.96	0.99
s_6	0.92	0.00	0.91
s_7	0.90	0.31	0.93
s_8	0.84	0.61	N/A

Table 5 Reliability coefficients report by attribute

Station attributes	Applicant reliabilities	Rater reliabilities	Attribute reliabilities
Communication	0.76	0.01	0.94
Critical thinking	0.74	0.00	0.85
Maturity	0.73	0.48	0.85
Empathy	0.67	0.07	0.63
Professionalism	0.66	0.73	0.99

applicants across these three attributes with a high degree of similarity. Station 6 presents the ideal conditions for reliability: high for applicants, low for raters, and high for items. In contrast, Station 3 is problematic. The reliability coefficient for items at station 3 is 0.07, suggesting that the three attributes (communication, critical thinking, and maturity) were not measured independently of each other. Therefore, it remains unclear if distinct attributes are being measured at this particular station.

Further analyses of the attributes and a summary of the reliability coefficients associated with attributes are presented in Table 5. Although there were eight attributes measured across the eight stations in the MMI, only attributes that appeared in more than one station were included in this analysis. The applicant reliability coefficients for the five included attributes ranged from 0.66 to 0.76; this is remarkably lower than the observed values across the stations, but still indicative of moderate heterogeneity among applicants. The low rater reliability coefficients for attributes such as communication and critical thinking suggest that the faculty and student raters were internally consistent in their understanding of these particular attributes. Meanwhile, others (i.e., maturity and professionalism) produced substantially higher reliability coefficients, which accounts for some of the variability in rating behaviour across attributes, and thus stations.

The most problematic feature of the attribute analysis was the large reliability coefficients produced by the attributes themselves. Recall from earlier that high reliability values are generally desired for items; however, in this case low reliability coefficients are regarded as better because the attributes should be consistent regardless of the station. As an example, the attribute reliability for professionalism is 0.99, which indicates high levels of inconsistency in how professionalism is viewed across stations.

Discussion

Previous research has suggested the MMI can be a valuable admissions selection tool (e.g., Eva et al. 2004; Reiter et al. 2007) given that it reduces the biases commonly associated with more traditional interview procedures. With the importance of health education, developing admission techniques that provide consistent and accurate results is crucial. Additionally, with the increase in student advocacy and challenges to negative admission decisions, defensible application selection procedures are a necessity. For such reasons, the MMI is becoming an increasing popular method of identifying qualified candidates for post-secondary education in health professions. The similarity of MMI to the commonly used OSCEs makes it an ideal method as it provides a similar assessment context used in several health education programs. An added feature of the MMI method is its ability to measure valuable information concerning applicants' non-cognitive personal characteristics, abilities, and skills on desirable nonmedical traits.

Nevertheless, there remains a need to continue to examine the psychometric properties of the MMI, including the overall reliability and the general consistency of raters' abilities to score applicants accurately across desired traits. Our findings suggest the MMI meets some of the requirements for psychometric quality; however, there are still ongoing issues that need to be addressed. It appears that the MMI is able to capture certain non-cognitive aspects of applicants, which provides a more comprehensive account of prospective medical students. Our analyses were also able to identify problematic applicants, raters, stations, and items. Nevertheless, what remains unclear is what non-cognitive attributes were being assessed. Some of the attributes may be more challenging to measure than others, and thus present themselves as being more erratic across the group of applicants. Furthermore, the raters were unable to distinguish amongst the different non-cognitive attributes. Such evidence suggests a holistic measurement process, or the measurement of a broader unidimensional construct across the twenty-one items used in this MMI. This construct may be better defined as "suitability for medical school" or "professionalism." According to Arnold (2002), approximately 50 % of medical schools develop their own criteria for measuring professionalism. Many institutions have reported using elements such as honesty, maturity, integrity, and other interpersonal skills to measure professionalism among medical applicants (Arnold 2002). The attempt to capture the essence of professionalism through the measurement of alternative constructs may also explain the noise we observed with the professionalism attribute in this MMI.

Results from the G theory analyses suggest that the majority of the observed variance in applicants' scores was due to the score differences across applicants—the interaction between applicants and stations—and other confounding unexplained sources of variance. Otherwise stated, the choice of stations matter. Moreover, since there was an applicant-station interaction, using the MMI instead of the single interview generated more variability between applicants, which provides further sources of information that could be valuable for admissions decisions. The subsequent D studies found that increased generalizability was best obtained by increasing the number of stations rather than the number of raters. As long as the additional stations measure unique dimensions with the same degree of reliability, the overall content validity of the assessment will increase.

While there was moderate rater consistency within stations, additional rater training can still play a role in reducing unwanted variation. A 9-point scale was established for scoring procedures; however, only three of the points were actually defined. Thus raters' notions of performance may be inconsistent using this limited information, as there were too many undefined values in the scale. Finally, there was a large residual error variance due to the design of the study, which confounds the effects among the facets of measurement.

The MFRM was able to produce logit measures and Fit Statistics identifying inconsistent raters, stations, and items. These statistics provide direction to make adjustments and correct for notable differences in order to ensure that all applicants are measured fairly across admissions' criteria. The psychometric results of the MMI using the Rasch model suggest that all of the non-cognitive attributes fit within a broader unidimensional construct rather than the individual non-cognitive attributes. The report of the probability curves generated by the 9-point rating scale also suggest that certain points on the rating scale could be collapsed, as the categories do not appear to be different enough to warrant separate categories. Given that the rating scale used to evaluate the medical school applicants was somewhat ambiguous from the beginning, with individual points not defined, adjustments to the rating scale or reducing its range may help endorse the non-cognitive attributes measured by the MMI.

While the average scores of the faculty and student raters were similar, the analysis of reliability coefficients for the raters revealed notable differences between the faculty and students across the eight stations. These results suggest an internal consistency issue, and providing sufficient training and education to both the faculty and students raters may resolve some of the inconsistencies among the raters. Without appropriate rater training, the raters may begin to impose their own values and beliefs about what constitutes a particular attribute (e.g., professionalism), which could lead to bias and unfair evaluations of applicants. Furthermore, adequate training makes raters aware of the commonly identified errors that can occur by providing information and teaching raters how to be more objective as they assess applicant behaviours. A more comprehensive form of rater training is needed to ensure consistency among the raters, both internally and externally. Such training would help to distinguish the criteria used to measure each of the intended attributes. The cost of training raters and the development of new stations should also be considered, as it can be substantial.

While the ratings across stations were similar among medical faculty and students' scores for this MMI, the arrangement used in the admissions procedure prevented us from determining the extent to which certain raters were more stringent or lenient than others. As a result, the measures only examined the extent of scores similarities across the stations. Subsequent research is required to measure individual variability of raters' scoring practices. One possibility could include the exploration of a fully-crossed design (e.g., using the same raters at each station), but it should be acknowledged that while a fully-crossed design would allow for a larger universe of generalization, it would also pose logistical challenges during administration.

Nevertheless, our results across the G study and Rasch analyses highlight the real potential of the MMI to support admissions decisions while also identifying potential challenges. Based on our analyses, the MMI can provide a consistent measure of applicants' suitability for admissions. At this time, we do not have sufficient evidence that the efforts to measure different non-cognitive attributes have been successful. What remains unclear is how difficult it is for each of the raters to provide distinct scores on each of the intended attributes. Hence it may be more worthwhile, and less taxing on the raters, to provide a more holistic score at each of the stations. Lastly, the choice of station matters in the use of the MMI to support admissions decisions. Given the station variability we observed, it is certainly possible that the use of different stations would have resulted in the selection of different applicants. Such a finding has important implications for the selection of MMI stations for subsequent use.

Appendix

See Table 6.

Table 6 List of items (attributes) for MMI stations

Item	Attribute	Where attribute was measured
1	Communication	Station 1
2	Critical thinking	Station 1
3	Maturity	Station 1
4	Communication	Station 2
5	Critical thinking	Station 2
6	Communication	Station 3
7	Critical thinking	Station 3
8	Maturity	Station 3
9	Effectiveness	Station 4
10	Empathy	Station 4
11	Professionalism	Station 4
12	Resolution	Station 4
13	Integrity	Station 5
14	Maturity	Station 5
15	Communication	Station 6
16	Empathy	Station 6
17	Critical thinking	Station 6
18	Critical thinking	Station 7
19	Professionalism	Station 7
20	Maturity	Station 7
21	Communication	Station 8

References

Arnold, L. (2002). Assessing professional behavior: Yesterday, today, and tomorrow. *Academic Medicine, 77*(6), 502–515.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (American College Testing Technical Bulletin No. 43). Iowa City, IA: ACT.

Dodson, M., Crotty, B., Prideaux, D., Carne, R., Ward, A., & de Leeuw, E. (2009). The multiple mini-interview: How long is long enough? *Medical Education, 43*(2), 168–174.

Eva, K. W., Reiter, H. I., Rosenfeld, J., & Norman, G. R. (2004). An admissions OSCE: the multiple mini-interview. *Medical Education, 38*, 314–326.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counselling psychology research. *Journal of Counseling Psychology, 45*(1), 30–45.

Harden, R. M., & Gleeson, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education, 13*(1), 41–54.

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

Linacre, J. M. (1995). Categorical misfit statistics. *Rasch Measurement Transactions, 9*(3), 450.

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

Linacre, J. M. (2011). *Facets computer program for many-facet Rasch measurement*, version 3.68.1. Beaverton, OR: Winsteps.com.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331–345.

Reiter, H. I., Eva, K. W., Rosenfeld, J., & Norman, G. R. (2007). Multiple mini-interview predicts for clinical clerkship performance? National licensure examination performance. *Medical Education, 41*(4), 378–384.

Roberts, C., Walton, M., Rothnie, I., Crossley, J., Lyon, P., Kumar, K., et al. (2008). Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Medical Education, 42*, 396–404.

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

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 81–124). Dordrecht: Elsevier.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.