The Ability of the Multiple Mini-Interview to Predict Preclerkship Performance in Medical School

KEVIN W. EVA, HAROLD I. REITER, JACK ROSENFELD, and GEOFFREY R. NORMAN

Abstract

Problem Statement and Background. One of the greatest challenges continuing to face medical educators is the development of an admissions protocol that provides valid information pertaining to the noncognitive qualities candidates possess. An innovative protocol, the Multiple Mini-Interview, has recently been shown to be feasible, acceptable, and reliable. This article presents a first assessment of the technique's validity.

Method. Forty five candidates to the Undergraduate MD program at McMaster University participated in an MMI in Spring 2002 and enrolled in the program the following autumn. Performance on this tool and on the traditional protocol was compared to performance on preclerkship evaluation exercises.

Results. The MMI was the best predictor of objective structured clinical examination performance and grade point average was the most consistent predictor of performance on multiple-choice question examinations of medical knowledge.

Conclusions. While further validity testing is required, the MMI appears better able to predict preclerkship performance relative to traditional tools designed to assess the noncognitive qualities of applicants.

Despite widespread recognition that noncognitive (i.e., personal) qualities are important contributors to the ability of students to become competent physicians,[1,2] evidence that selection tools can accurately identify candidates in possession of the desired characteristics has remained elusive.[3] To gain admission to almost any professional school in North America, programs require candidates to complete a personal interview during which an attempt is made to assess each candidate's interpersonal skills, motivation, and problem exploration.[4] There has been tremendous variability in the reliability observed within studies of the interview process, some of which can be accounted for by the finding that more structured interviews (i.e., interviews that utilize standardized questions) tend to elicit more reliable judgments.[5] However, the vast majority of these studies have been interrater reliability studies, their results indicating simply that different individuals can sometimes agree on the strength of a performance within a single interview.

To gain useful information about the candidate, however, some reassurance is required that the admissions tool provides information that is generalizable beyond that particular interview. Several studies investigating the conduct of multiple encounters have demonstrated a poor correlation of candidate performance across interviews.[6–8] With this as background, it should come as no surprise that interview scores have repeatedly failed tests of validity. Twenty-five years ago, Mann[9] was able to show that interview ratings depend more on the interviewer than on the candidate. More recently, it has been shown that traditional interviews do not predict performance on objective structured clinical examinations (OSCE) either within medical school[10] or during licensure,[11] despite the fact that OSCEs tend to be designed, in part, to specifically address some of the noncognitive skills that interviews are expected to assess (e.g., communication skills, problem exploration etc.,).[12]

Adopting a lesson from the evaluation community—the need to collect multiple observations whenever context specificity is a concern—Eva et al.[8] have developed a Multiple Mini-Interview (MMI) process in an attempt to dilute the impact of individual examiners and allow for a more generalizable aggregate performance rating to be assigned to candidates. Modeled after the OSCE, the MMI consists of a series of short interviews with multiple examiners. Each station is assigned a scenario pertaining to ethical issues, communication skills, collaborative abilities, or some other noncognitive quality. Greater detail regarding the methodology can be found elsewhere, but it should be noted that a series of studies have shown the MMI to be feasible, acceptable to both candidates and examiners, and reliable (G = .65–.81).[8,13] This article presents the first attempt to assess the validity of this new admissions protocol by examining the relationship between preclerkship performance, the MMI, and the traditional admissions protocol used by the Undergraduate MD program at McMaster University.

Method

Participants

Forty-five of the 117 Undergraduate MD program candidates who participated in the MMI in Spring 2002 enrolled in the program the subsequent autumn. These 45 students constitute 32.6% of the class of 2005. At the time of this study, these students had completed the preclerkship component of the curriculum.

Admissions Tools

The admissions cycle for the class of 2005 began in October 2001 at which time all candidates (n = 3,027) submitted an application package to the Ontario Medical School Application Service, a central repository that collects applications for all five Ontario medical schools. The portion of the application used at McMaster consisted primarily of student transcripts (from which grade point average [GPA] was calculated) and an autobiographical submission (ABS). The ABS required candidates to provide short essay-type responses to a series of 15 questions indicating, for example, their past experiences and their reasons for wanting to become a physician. Each ABS was read by three raters (one faculty member, one community member, and one student) and marked using a seven-point scale. GPA and ABS were used as a screen of candidates' cognitive and noncognitive qualities, respectively. Each was weighted equally and used to determine the top 384 candidates, who were invited to interview in March/April of 2002.

During interview weekend, invited candidates proceeded through a series of tasks. The first, a traditional personal interview, took place with three examiners (again, one representative from each stakeholder group) and lasted 30 minutes. After the interview, examiners were provided 30 minutes to gather their thoughts and independently rate the candidate, again using a seven-point scale. Next, candidates were gathered into groups of six, provided with a pair of problems, and given 30 minutes to demonstrate their ability to work through these problems as a group. This exercise is referred to as a simulated tutorial.[14] Three examiners observed the interac-

Table 1. Uncorrected Correlations between Admissions Tools and In-Program Objective Structured Clinical Examinations (OSCE) and Personal Progress Inventories (PPI)

Admissions Tool	Inaugural OSCE	Preclerkship OSCE	PPI #1	PPI #2	PPI #3	PPI #4
Autobiographical submission	.05	−.08	.29*	.25*	.17	.15
Grade point average	.10	−.04	.15	.20	.26*	.28*
Multiple Mini-Interview	.32*	.23	.09	.12	.27*	.24*
Personal interview	.08	.11	.12	.01	.07	−.12
Simulated tutorial	−.04	−.20	−.01	−.25*	−.11	−.11

*$p < .10$.

tion from behind a one-way mirror and rated each candidate's performance. Both of these tools were designed with the intent of measuring candidates' noncognitive qualities. Each student's performance was considered as a whole, and a collation committee made the final judgment regarding who would be offered admission to the program.

Finally, separate from the formal admission process, each of the 384 candidates invited to interview were also invited to participate in a research study pertaining to a new admissions tool—the MMI. The first 120 who could be booked into prearranged study sessions were enrolled to participate. Each candidate participated in a ten-station MMI, during which each station lasted eight minutes and was rated by a single examiner using a seven-point scale. The question asked of examiners was "Please rate the applicant's **overall performance** on this station." The Points 1, 3, 5, and 7 were anchored with the adjectives "Unsatisfactory," "Borderline," "Satisfactory," and "Outstanding." Like the traditional interview and the simulated tutorial, the MMI was also designed to provide information regarding candidates' noncognitive qualities. Greater detail regarding the design and procedure is provided elsewhere.[8]

Preclerkship Evaluation Tools

Since the early 1990s, McMaster's Undergraduate MD program has required all students to complete a Personal Progress Inventory (PPI) three times per year during their tenure in the program. The PPI is a multiple-choice question examination that consists of 180 questions. It is intended to provide a measure of competency in cognitive domains, broadly testing medical knowledge. Reliability and validity analyses suggest that it accomplishes this goal. Test–retest reliability has been shown to be .70 over successive test intervals, and there is a significant correlation between PPI performance and subsequent performance on the national licensing examination ($r = .62$ for PPIs written later in training).[15] At the time of this study, the class of 2005 had completed four PPIs.

In addition, McMaster students participate in an OSCE each year in the program. The OSCE, as it is used at McMaster, consists of ten stations, each of which requires students to perform a physical maneuver, take a history, or provide standardized patients with information. Faculty observers assess each candidate's performance using a seven-point global rating scale. The goal of this tool, in part, is to provide information pertaining to an applicant's noncognitive competencies, including communication skills and problem solving ability. At the time of this study, the class of 2005 had completed an inaugural OSCE in Spring 2003 and a preclerkship OSCE in late Autumn 2003.

Finally, during each of the program's four preclerkship units, students encounter the majority of the curriculum in small problem-based tutorials. At the end of each tutorial, tutors are responsible for stimulating discussion around the dynamics of the group interaction and the performance (including professional behaviors) of group members. These discussions are then summarized on a qualitative end-of-unit evaluation form and accompanied by a rating of Satisfactory, Provisional Satisfactory, or Unsatisfactory.

We examined the correlation between admissions measures and in-course measures using Pearson's correlation coefficient and regression analyses. Because of the small sample size and an expectation (based on past research[11]) of attenuated correlations less than .3, the critical alpha level was set at .1.

Results

While sampling from enrollees inevitably results in differences between the sample and the population of all candidates, all of the evidence we collected suggests that the 45 students who completed the MMI were no different than the other students enrolled in the MD program; PPI and OSCE scores received by students in our sample were within 1% of the scores received by students who did not complete the MMI; each comparison was nonsignificant.

The internal consistency (Cronbach's alpha) of the inaugural OSCE and the preclerkship OSCE was found to be .57 and .69, respectively. The intertest reliability of the PPI was 0.85.

Table 1 illustrates the correlation between each admissions tool and the various preclerkship evaluation tools. Disattenuating each correlation for the imperfect internal consistency of the OSCE improves each of the correlations in Table 1 by approximately 50%. In predicting OSCE performance, the MMI clearly outperformed each of the traditional admissions protocols. Inclusion of all five admissions tools in a regression analysis as predictors of mean OSCE performance revealed that only the MMI was statistically predictive ($\beta = .44$, $p < .01$), as illustrated in Table 2.

Table 2 also reveals that GPA and ABS were observed to be the most consistent predictors of mean PPI performance when entered into a regression analysis. None of the other admissions tools were statistically predictive of PPI performance. Despite equivalent predictive ability, the nature of the predictions provided by ABS and GPA were qualitatively quite different, as illustrated in Table 1. ABS scores had a relatively strong relationship with PPI performance early in the program, but these correlations rapidly declined. In contrast, GPA did not predict performance on the first PPI, but steadily improved in its ability to predict later PPI performances. Excluding ten nonscience students from the analyses results in increased correlations between GPA and PPI (between .39 and .46

Table 2. Standardized Coefficients (β) Indicating the Ability of Each Admissions Tool to Predict Mean Objective Structured Clinical Examination (OSCE) and Mean Personal Progress Inventory (PPI) Performance

Admissions Tool	OSCE	PPI
Autobiographical submission	−.12	.45*
Grade point average	.05	.54*
Multiple Mini-Interview	.44*	.26
Personal interview	.06	.01
Simulated tutorial	−.23	−.15

*$p < .05$.

for each PPI) and decreased correlations between ABS and PPI (between .01 and .05 for each PPI).

Finally, at the time of this study, one of the 45 students in the sample had received an unsatisfactory rating on tutorial performance for professional behavior reasons. Such a judgment is made very infrequently. While one student does not provide enough information to allow conclusive judgments, it is interesting to note this student's performance on the various admissions tools. When compared to the 45 students included in the sample, the MMI ranked this student lowest (in the 15th percentile) relative to the other four admissions tools, indicating that the MMI was most likely to have excluded this student when admissions decisions were reached. In contrast, this student was ranked relatively high by the simulated tutorial (76th percentile), grade point average (76th percentile), and the personal interview (48th percentile), indicating a higher likelihood of being offered admission based on these tools alone. The student was in the 17th percentile based on ABS scores.

Discussion

The results of this study support the hypothesis that the MMI provides a more valid indication of a candidate's noncognitive characteristics than do more traditional admissions tools. The lack of correlation between the traditional personal interview and OSCEs replicates the work of Basco et al.[10] Furthermore, the finding that GPA most consistently predicts performance on cognitive outcome measures replicates the findings of Kulatunga-Moruzi and Norman[11] almost perfectly; the correlation they observed between undergraduate GPA and national licensing examination performance was .33 and the correlations they observed between the personal interview and simulated tutorial and the national licensing examination were $r = 0.03$ and 0.01, respectively.

It is interesting to note that GPA was not statistically predictive of performance on the very first PPI, written one month after entry into the medical program, but that the strength of the association appears to increase over time. The most plausible explanation, given the change in correlations observed upon excluding students who entered the program without an undergraduate science degree, is that these individuals diluted the magnitude of the relationship. High grades in an arts program cannot be expected to translate to high grades on a test of medical knowledge less than one month into medical school. Continued monitoring of PPI performance during clerkship and collection of larger samples in subsequent cohorts will assist in determining the extent to which background specific findings might have arisen due to chance.

A potential explanation of the finding that the MMI was most predictive of OSCE performance is that the MMI is the only tool that did not suffer from restriction of range as a result of its being the only tool that did not contribute to the final admission decision for the class of 2005. This hypothesis was not supported by the data. For the 45 individuals included in this study, the coefficient of variation (standard deviation divided by mean), a measure of the amount of variability in the data, was lower for the MMI (.13) than for either of the other tools used during the interview weekend (personal interview = .19, simulated tutorial = 0.22). Despite this finding, none of the traditional tools were able to consistently predict performance on either the PPI or the OSCE. If anything, the correlation is negative in some instances.

The most disconcerting finding in this study was that the simulated tutorial resulted in a high ranking for a student who was eventually identified as a concern regarding professional behaviors in tutorial. Grade point average resulted in an equally poor ranking, but the simulated tutorial has been designed and used with the express purpose of attempting to identify and select for the qualities that enable productive collaboration within a tutorial setting. Again, it should be noted that this analysis focuses on only one student, but a more systematic study of students achieving this degree of concern is unlikely given the low frequency with which tutors have concerns grave enough to warrant official documentation; an unsatisfactory rating results in the student being removed from the program until such time as remediation has been completed. Alternative strategies to assess this aspect of performance will be implemented in the future.

In addition, further validity testing is underway to determine precisely which noncognitive characteristics are being captured by the MMI and, more specifically, whether or not specific stations are able to capture the precise noncognitive characteristics for which they were designed. One of the more compelling aspects of the MMI is the potential for individual institutions to tailor the stations toward selection of the characteristics that are most valued within the local context.[16] To ensure this is possible, discriminant validity testing will be completed. In addition, a more longitudinal assessment of the predictive value of the MMI is required. The short-term duration of follow-up and limited sample size are, in fact, the primary limitations of this study. Nonetheless, this preliminary analysis of the ability of the MMI to predict performance within medical school bodes well, especially when compared to the predictive value of traditional admissions tools that continue to be used worldwide.

This project was funded (in part) by a National Board of Medical Examiners (NBME®) Edward J. Stemmler, MD Medical Education Research Fund grant. The project does not necessarily reflect NBME policy, and NBME support provides no official endorsement.

Correspondence: Kevin W. Eva, PhD, Department of Clinical Epidemiology and Biostatistics, Programme for Educational Research and Development, Room 101, T-13, McMaster University, Hamilton, Ontario, L8S 4K1, Canada; e-mail: ⟨evakw@mcmaster.ca⟩.

References

1. Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. Assessing personal qualities in medical school admissions. Acad Med. 2003;78:313–21.
2. Royal College of Physicians and Surgeons of Canada. CanMEDs 2000 Project: Skills for the new millennium. Report of the societal needs working group, 1996 ⟨http://rcpsc.medical.org/canmeds/index.php⟩. Accessed 15 June 2004.
3. Salvatori P. Reliability and validity of admissions tools used to select students for the health professions. Adv Health Sci Educ. 2001;6:159–75.
4. Morris JG. The value and role of the interview in the student admissions process: a review. Med Teach. 1999;21:473–81.
5. Edwards JC, Johnson EK, Molidor JB. The interview in the admission process. Acad Med. 1990;65:167–75.
6. Turnbull J, Danoff D, Norman GR. Content specificity and oral examinations. Med Educ. 1996;30:56–9.
7. Kreiter CD, Yin P, Solow C, Brennan RL. Investigating the reliability of the medical school admissions interview. Adv Health Sci Educ. 2004;9:147–51.
8. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the Multiple Mini-Interview. Med Educ. 2004;38:314–26.
9. Mann WC. Interviewer scoring differences in student selection interviews. Am J Occup Ther. 1979;33:235–9.
10. Basco WT, Gilbert GE, Chessman AW, Blue AV. The ability of a medical school admission process to predict clinical performance and patients' satisfaction. Acad Med. 2000;75:743–7.
11. Kulatunga-Moruzi C, Norman GR. Validity of admissions measures in predicting performance outcomes: the contribution of cognitive and non-cognitive dimensions. Teach Learn Med. 2002;14:34–42.
12. Reznick RK, Blackmore D, Cohen R, et al. An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality. Acad Med. 1993;68:S4–6.
13. Eva KW, Reiter HI, Rosenfeld J, Norman GR. The relationship between interviewer characteristics and ratings assigned during a Multiple Mini-Interview. Acad Med. 2004;79:602–9.
14. Ferrier BM, McAuley RG, Roberts RS. Selection of medical students at McMaster University. J R Coll Phys Lond. 1978;12:365–78.
15. Blake JM, Norman GR, Smith EKM. Report card from McMaster: student evaluation at a problem-based medical school. Lancet. 1995;345:899–902.
16. Reiter HI, Eva KW. Reflecting the relative values of community, faculty, and students in admissions tools for medical school. Teach Learn Med. In press.