The multiple mini-interview in the UK context: 3 years of experience at Dundee

JON DOWELL, BONNIE LYNCH, HETTIE TILL, BEN KUMWENDA & ADRIAN HUSBANDS
University of Dundee, UK

Abstract

Background: The Multiple Mini-Interview (MMI) is a new interview process that Dundee Medical School has recently adopted to assess entrants into its undergraduate medicine course. This involves an ‘Objective Structured Clinical Examination’ like rotational approach in which candidates are assessed on specific attributes at a number of stations.

Aims: To present methodological, questionnaire and psychometric data on the transitional process from traditional interviews to MMIs over a 3-year period and discuss the implications for those considering making this transition.

Methods: To facilitate the transition, a four-station MMI was piloted in 2007. Success encouraged consideration of desirable attributes which were used to develop a full 10-station process which was implemented in 2009 with assessors being recruited from staff, students and simulated patients. A questionnaire was administered to all assessors and candidates who participated in the 2009 MMIs. Cronbach’s alpha and Pearson’s $r$ and analysis of variances were used to determine the MMI’s psychometric properties. Multi-faceted Rasch modelling (MFRM) was modelled to control for assessor leniency/stringency and the impact of using ‘fair scores’ determined. Analysis was conducted using SPSS 17 and FACETS 3.65.0.

Results: The questionnaire confirmed that the process was acceptable to all parties. Cronbach’s alpha reliability was satisfactory and consistent. Graduates/mature candidates outperformed UK school-leavers and overseas candidates. Using MFRM fair scores would change the selection outcome of 6.2% and 9.6% of candidates in 2009 and 2010, respectively. Students were less lenient, made more use of the full range of the rating scales and were just as reliable as staff.

Conclusions: The strategy of generating institutional support through staged introduction proved effective. The MMI in Dundee was shown to be feasible and displayed sound psychometric properties. Student assessors appeared to perform at least as well as staff. Despite a considerable intellectual and logistical challenge MMIs were successfully introduced and deemed worthwhile.

Introduction

Selection of candidates for medical school in the UK, as well as elsewhere, is typically based on several criteria: grades attained (or predicted) from school or college; the personal statement; and letters from professional references, such as head teachers. More recently in the UK, aptitude testing has been introduced (such as UK Clinical Aptitude Test http://www.ukcat.ac.uk/home/) and in most schools, a personal interview is also conducted. The interview has traditionally been used for the purposes of information gathering, decision making, verification and recruitment (Edwards et al. 1990). It is widely believed that the interview makes a unique contribution to the admissions process by gathering information on a candidate’s interpersonal skills vital to performance in the clinical setting (Meredith 1982).

Although relatively little has been published about medical school interviews, research across academic and applied domains consistently suggests that traditional interviews suffer from a lack of psychometric robustness. Challenges to their reliability and validity include interviewers’ proclivity for strictness or laxity (Mann 1979), ineffective training for interviewers (Johnson & Edwards 1991), lack of adequate structure and/or lack of an explicitly defined scoring rubric (Nayer 1992). Given these problems, the low reliability associated with many medical school admissions interviews (Kreiter et al. 2004) is not surprising. Around the world, medical schools’ desire to apply the similarly rigorous standards of evidence to their admissions processes that they
have typically applied to their clinical practice have resulted in the transitional phase being replaced by the recently developed Multiple Mini-Interview (MMI).

The MMI is conceptually based on the Objective Structured Clinical Examination (OSCE), where clinical competencies are repeatedly and rapidly sampled to provide a more accurate measure of ability in selected areas (Eva et al. 2004). Candidates rotate through a number of stations responding to questions, or discussing scenarios while being rated by one assessor using standardised scoring descriptors (Lemay et al. 2007).

The psychometric properties of the MMI have been investigated by North American and Australian medical schools who have consistently found that MMIs are more reliable and valid assessment tools than traditional interviews (Eva et al. 2004; Lemay et al. 2007; Reiter et al. 2007). The response to the introduction of MMIs has been positive from candidates, assessors and other stakeholders alike (Brownell et al. 2007; Humphrey et al. 2008; Razack et al. 2009).

Although MMIs are clearly a major advance on traditional interviews, there are psychometric issues that need to be addressed. As with OSCEs, the impact of assessor leniency/stringency as a source of error is an ongoing concern. One way in which this could be ameliorated is through adjusting candidates’ scores using Multi-faceted Rasch modelling (MFRM). Roberts et al. (2010) found that using MFRM fair scores would result in 11.5% of candidates seeing a significant change to their ranking for selection into medical school. Similarly McManus et al. (2006) found that in the case of a postgraduate licensing OSCE, the outcome for 4.1% of candidates would change after the adjustment. The impact of using these adjusted ‘fair scores’ could therefore be of significant benefit to those involved in final selection of candidates into medical schools (Roberts et al. 2010).

To date, there has been one study examining the introduction and performance of MMIs in the UK setting. O’Brien et al. (2011) ran traditional interviews alongside MMIs and found them to be reliable, feasible and acceptable to both applicants and interviewers in selecting undergraduate and graduate streams. However, a relatively small number of applicants undertook these MMIs and no UK-based study has detailed the MMI implementation process. Further psychometric analyses in a UK population exploring key aspects of the process are still required.

The purpose of this study is to describe Dundee medical school’s progression from a traditional interview to a full-scale MMI, psychometric properties, views of applicants and assessors and to identify areas of further research. It also raises issues those planning to introduce MMIs may wish to consider such as generating institutional support, the value of using students as assessors and the potential value of adjusting for examiner leniency and stringency.

Methods

Progression from traditional interview to MMIs

Before implementing a full-scale MMI in 2009, we felt that a transitional phase was needed, in which elements of the new system could be integrated with the existing one. We chose this approach for two reasons: stakeholders would likely accept and support the new system, if it did not appear to be a large, abrupt change to the existing system; and the materials development, logistics and other resource demands of the full-scale MMI would be very difficult to meet within a single admissions cycle. In 2007, we therefore settled upon a pilot MMI that was essentially a hybrid of the old interview content using a new rotational approach. Other changes were the introduction of student-assessors and formalised training.

The piloted system comprised four 10-min stations, three of which were simply traditional one-to-one interviews, each employing questions adapted directly from the previous format. The fourth station was a new student run interactive assessment in which candidates were observed interacting with a role player to complete a task. This station focused on teamwork and empathy, attributes widely considered important for success in medicine that were not being assessed explicitly by the old system. All assessors received 15–20 min of general training and 15 min of station-specific training on the day of their first session. Assessors evaluated candidates on three attributes per station, using 5-point Likert scales. A candidate could therefore receive a maximum score of 15 points in each station, and a total score of 60 points.

In the 2007–2008 admissions cycle, we performed 473 interviews. This represented a 25% increase from the previous year, but was resource-neutral for staff because students ran one station. Systematically gathered informal feedback suggested that assessors and candidates were satisfied with the process.

Cronbach’s alpha computed from the total scores for the four MMI stations was 0.66 with a range of station-total score correlations between 0.259 and 0.387. Overall, these results provided reassurance that the format was functional with the brief training approach employed and convinced faculty that we were able to create and run a reliable MMI.

Introduction of 10 station MMI

Introducing an MMI introduced a number of challenges such as recruiting assessors, acquiring appropriate space, orchestrating the process, generating content and maintaining data integrity.

After examining the research done by Eva et al. (2004), observing the MMI in McMaster, and reassured by the pilot data, we determined that 10 stations, with one assessor per station, would be the optimal balance for our purposes. Development of specific attributes to be assessed followed a similar process to that reported elsewhere (Reiter & Eva 2005). A superset of possible attributes was generated via a literature search that included journals and standards documents published by relevant governing bodies for the practice of medicine in the UK. This superset was presented to the medical school’s admissions committee for discussion and ranking. The primary objective was to define a set of core attributes that the committee deemed important, consistent with the school’s mission statement, and appropriate to the live nature of the interview (i.e. not duplicating other existing selection components). This process resulted in a set of six
attributes or ‘domains’:

- Interpersonal skills and communication (including empathy)
- Logical reasoning and critical thinking
- Moral and ethical reasoning
- Motivation and preparation to study medicine
- Teamwork and leadership
- Honesty and integrity

Specific content and scoring rubrics for each station were generated and refined by three staff members: a cognitive psychologist and lecturer in medical education, director of undergraduate admissions (and a general practitioner) and a doctoral student involved in research on an MMI-type assessment. There were seven stations in common between the 2009 and 2010 MMIs, with three being unique. Station names with brief descriptions are provided in Table 1.

Each station was designed to elicit information about three of the six domains and with each score on a 5-point Likert scale. The sum of the three attribute scores at a single station constituted the station score. The sum of the station scores constituted the overall MMI score (total 30–150). The inclusion of different numbers of scores for different attributes produced a weighting scheme that accorded approximately with the relative importance of the attributes, as determined by the rankings from the admissions committee (with the exception of integrity which was recognised as the most challenging to assess).

Each station also included a ‘red flag’ option which gave assessors an opportunity to express severe concerns about a candidate’s suitability. Assessors who tick the ‘red flag’ box were required to explain the reason. Examples included candidate rudeness to an assessor or another candidate, bizarre statements or mannerisms suggesting they are unable to conduct themselves in a professional manner. The intention was that two independent red flags would exclude an applicant if deemed relevant after review by the University’s Medical Admissions Committee. Thus, no one assessor’s view could prevent admission.

Assessors were recruited from three sources, primarily via email appeals: clinical and non-clinical staff, medical students and simulated patients. They were asked to contribute at least two half day each, as this was required for Rasch analysis which we were eager to consider.

On their first day, all assessors received general training about the MMI, rationale for its use and logistics of running the stations. This training lasted approximately 30 min. Thereafter, each assessor spent 25–30 min in their assigned station, reading station-specific training materials and rehearsing where necessary. The authors answered questions and monitored performance of the stations by direct observation. Two of the 10 stations involved the use of standardised roles, these student actors received a half days training in advance. Actors and other role players did not participate in scoring the candidates.

Candidates were oriented to the process in several ways. The admissions website provided general information about the format and rationale. The letter sent to candidates to invite them for interview reiterated this information. On the day of their interview, candidates received a brief reminder and more details about the logistics of rotation through the stations.

Each station was 7 min in duration (a pragmatic maximum achievable), including time to rotate and read the station instructions. On any given day, a maximum of two full circuits (10 in each) was run in the morning and two concurrent sets of circuits (i.e. four full circuits) were run in the afternoon; therefore, a total of 60 candidates could be interviewed each day. After the interviews, candidates were invited to take an informal tour of the medical school.

### Analysis of scores

Parametric statistical techniques were used as the station-total scores had an approximately normal distribution. To determine whether the MMI was a fair assessment we conducted analysis of variances (ANOVAs) with candidate qualification groups as independent variables and station scores and overall MMI scores as dependent variables. Qualification groups were comprised of UK School-leaver (A-levels and Scottish Highers), Graduate/Mature (mature or with degree-level qualifications), European Union (EU qualifications) and Overseas (qualifications from outside Britain and the EU). ANOVA post hoc comparisons used the Fisher’s least significant difference test. To determine the relationship between the MMI and pre-other pre-interview admissions measures

<table>
<thead>
<tr>
<th>Station</th>
<th>Description</th>
<th>Type</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Counselling interaction with a distressed fellow student</td>
<td>Interactive</td>
<td>2009 and 2010</td>
</tr>
<tr>
<td>2</td>
<td>Candidate plays the role of a peer advisor assessing a student who has admitted wrongdoing</td>
<td>Interactive</td>
<td>2009 and 2010</td>
</tr>
<tr>
<td>3</td>
<td>Candidate is given a complex card sorting task to accomplish within a short time</td>
<td>Interactive</td>
<td>2009 and 2010</td>
</tr>
<tr>
<td>4</td>
<td>Candidate and helper complete a puzzle task</td>
<td>Interactive</td>
<td>2009 and 2010</td>
</tr>
<tr>
<td>5</td>
<td>Assessor explores experiences and motivation to apply to study medicine</td>
<td>One-to-One</td>
<td>2009 and 2010</td>
</tr>
<tr>
<td>6</td>
<td>Discussion of expectations of career as a doctor</td>
<td>One-to-One</td>
<td>2009 and 2010</td>
</tr>
<tr>
<td>7</td>
<td>Discussion of topical issues on the provision of medical care</td>
<td>One-to-One</td>
<td>2009 and 2010</td>
</tr>
<tr>
<td>8</td>
<td>Discussion about learning styles</td>
<td>One-to-one</td>
<td>2009</td>
</tr>
<tr>
<td>9</td>
<td>Discussion about a controversial ethical issue involving a non-medical controversy</td>
<td>One-to-one</td>
<td>2009</td>
</tr>
<tr>
<td>10</td>
<td>Candidate explores a hypothetical moral dilemma</td>
<td>One-to-one</td>
<td>2009</td>
</tr>
<tr>
<td>11</td>
<td>Candidate interacts with an actor who has difficulty managing their medications</td>
<td>Interactive</td>
<td>2010</td>
</tr>
<tr>
<td>12</td>
<td>Discussion about a controversial ethical issue in medicine</td>
<td>One-to-one</td>
<td>2010</td>
</tr>
<tr>
<td>13</td>
<td>Candidate explores a hypothetical moral dilemma</td>
<td>One-to-one</td>
<td>2010</td>
</tr>
</tbody>
</table>
For personal use only.

For personal use only.

For personal use only.

For personal use only.

J. Dowell et al.

Table 2. MMI candidate demographics.

<table>
<thead>
<tr>
<th>Applicant group</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scottish higher</td>
<td>284</td>
<td>263</td>
</tr>
<tr>
<td>A-level</td>
<td>89</td>
<td>113</td>
</tr>
<tr>
<td>EU</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>OS</td>
<td>23</td>
<td>32</td>
</tr>
<tr>
<td>Graduate/mature</td>
<td>32</td>
<td>40</td>
</tr>
<tr>
<td>Assessors</td>
<td>91</td>
<td>72</td>
</tr>
<tr>
<td>Students</td>
<td>58</td>
<td>61</td>
</tr>
<tr>
<td>SPs</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td><strong>284</strong></td>
<td><strong>263</strong></td>
</tr>
<tr>
<td><strong>Mean age at application</strong></td>
<td><strong>18.5</strong></td>
<td><strong>18.5</strong></td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td><strong>254</strong></td>
<td><strong>280</strong></td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td><strong>198</strong></td>
<td><strong>197</strong></td>
</tr>
<tr>
<td><strong>%</strong></td>
<td><strong>43.8</strong></td>
<td><strong>41.3</strong></td>
</tr>
<tr>
<td><strong>%</strong></td>
<td><strong>56.2</strong></td>
<td><strong>58.7</strong></td>
</tr>
</tbody>
</table>

Table 3. MMI mean scores.

<table>
<thead>
<tr>
<th></th>
<th>2009 Mean (SD)</th>
<th>Score range</th>
<th>2010 Mean (SD)</th>
<th>Score range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>104.8 (11.8)</td>
<td>70–141</td>
<td>107.5 (12.1)</td>
<td>69–135</td>
</tr>
<tr>
<td>Male</td>
<td>104.0 (12.3)</td>
<td>70–130</td>
<td>105.3 (13.1)</td>
<td>63–133</td>
</tr>
<tr>
<td>Female</td>
<td>105.4 (11.3)</td>
<td>70–141</td>
<td>109.1 (11.1)</td>
<td>77–135</td>
</tr>
</tbody>
</table>

Table 2. MMI candidate demographics.

Table 3. MMI mean scores.

Pearson’s $r$ was used to correlate these with MMI scores. Pre-interview measures consisted of quantitative judgements of a candidate’s academic achievement, medical work experience and non-academic achievements (both derived from the personal statement), references and total UKCAT score (in quintiles). To investigate the MMI’s internal consistency reliability, Cronbach’s alpha was used. Analysis was performed with SPSS 17.0 for these statistical calculations.

Candidate and assessor views

To assess the MMI’s face validity and other aspects of its acceptance by assessors and candidates, we developed separate questionnaires for each group and conducted an online survey in 2009. This survey of all candidates (324/433) and assessors (116/157) was conducted shortly after the 2009 MMI (before decisions known to candidates). It utilised a 5-point Likert scale to attitudinal statements regarding the perceived fairness and validity of the overall process and the perceived validity and challenge of each station (for assessors, it also included ease of scoring). Candidates who already had experienced a traditional style interview elsewhere were asked to express their preference for Dundee’s MMI, and whether this would influence their choice of a medical school.

Multi-faceted Rasch modelling

A three-facet Rasch model was run on all MMI data to separate the effects of candidate ability, assessor leniency/stringency and the marking scale. This technique is an application of Item-Response Theory and has the capability to both identify these sources of error and partially adjust for them in a way that is fairer on candidates (Roberts et al. 2010). By calculating estimates of different ‘facets’, namely candidate ability, interviewer stringency/leniency and question difficulty a so-called ‘fair average score’ can be computed that reflects candidates’ abilities more accurately. The ‘fair score’ is the score that would have been obtained if all the measures of the other elements on all other facets had been located at the average

Results

Demographics

Results presented here are for 2009 (2008–2009 cycle) and 2010 (2009–2010 cycle) interview data; so the consistency of performance can be considered. Table 2 shows the breakdown of candidate demographics. A total of 452 and 477 candidates were interviewed in 2009 and 2010, respectively. Similar proportions of males, females and school leaving qualification groups (predominantly Scottish school-leavers) were interviewed in each year, reflecting the local applicant pool.

Analysis of scores

The three domain sub-scores gave candidates a possible total of 15 marks per station and a maximum score of 150 for the MMI. Descriptive statistics for total MMI scores are given in Table 3. Mean scores were not significantly different across the days of the MMIs or, indeed years and this suggests that there was no appreciable drift in scores due to potential leakage of station material.

A one-way ANOVA showed that there were significant differences between the graduate/mature, UK school-leaver and overseas candidate groups in both years, 2009: $F(2) = 7.75$, $p < 0.01$; 2010: $F(2) = 11.23$, $p < 0.01$. Post hoc comparisons showed that graduate/mature candidates achieved significantly higher mean scores than all other groups of candidates. Error bars with 95% confidence intervals for candidate groups are shown in Figure 1 and provides a graphical representation of the range of values within which the means of the populations may lie.

Cronbach’s alpha was computed from the total scores obtained for the 10 MMI stations in 2009 and 2010 and was 0.70 and 0.69, respectively. The range of correlations between station scores for 2009 and 2010 were 0.057 to 0.363 and −0.061 to 0.308, respectively. One station had a negative effect on the reliability in both years. This teamwork station was similar in each year with the scoring process modified before being run again in 2010. The analysis shows it requires further consideration.

Reliability was also calculated for each domain for 2009 and 2010. Where there was more than one related domain score in each station, the total score per domain within a station was
used (i.e. interpersonal communication and empathy scores). Domain reliability statistics are provided in Table 4.

The reliability of interactive (mostly student scored) and one-to-one (exclusively staff scored) stations was analysed for both years. Comparing these station types provides insight into the scoring behaviours of the assessor groups. Table 5 shows descriptive data, reliabilities of each station type and their correlations with the overall MMI score.

Table 6 shows descriptive statistics on the number of red flags given by assessors in 2009 and 2010. Three of the 660 ranked high enough to be considered for an offer were rejected after consideration of the qualitative comments accompanying a concern expressed by an assessor.

**Correlation of the MMI with other measures.** MMI scores were correlated with pre-interview selection measures to assess the overlap between the different elements of the process. As expected correlations were small (range from $-0.075$ to $0.123$ in 2009 and $-0.077$ and $0.067$ in 2010) and none reached statistical significance.

**Candidate and assessor views**

Questionnaires were returned and analysed from 324 (75%) of interviewees and 116 (58%) of assessors. An overwhelming
majority (94% and 90%, respectively) selected Agree or Strongly Agree in response to the question of whether the MMI was fair. A sizable minority of interviewees (33%) felt that the MMI was more stressful than the traditional interview; yet, 98% indicated that the use of the MMI either had no effect or actually enhanced their view of Dundee medical school. A large majority (74%) of those already interviewed at another medical school stated a preference for the MMI. A smaller minority (60%) said that they would recommend Dundee to a friend because of the MMI.

In total, 116 (116/157, representing 88% of student assessors and 64% of staff assessors) responded to the assessor questionnaire. A large majority felt that the MMI posed ‘very high’ or ‘moderate’ levels of stress for interviewees. Yet, an overwhelming majority (91%) also rated the fairness of the MMI as ‘very high’ or ‘moderate,’ and a similar percentage (94%) expressed very high or moderate willingness to participate in the process again. When asked about the ease of staffing the 10 stations, 5 of the stations received ratings of ‘very easy’ or ‘moderate’ from least 75% of respondents. Three stations received particularly low ratings (<50% of respondents offering the ‘very easy’ or ‘moderate’ rating). When asked how well the stations achieved what they set out to do, seven stations received ratings of ‘very well’ or ‘moderately well’ from at least 75% of respondents. One received concerning ratings (<47% of respondents assigning ‘very well’ or ‘moderately well’ ratings) and free-text comments suggested there was difficulty understanding the relevance of this critical thinking station which was the only one that involved an entirely non-medical controversy (fair trade and air transport of fruit from developing countries). Free-text comments suggested that two stations suffered from a lack of time to perform the assigned task. Consistent with these findings, a sizable minority (23%) expressed a need for more station-specific training.

Results – MFRM and fair scores

FACETS confirmed that sufficient linkage had been achieved between sessions/stations and assessors to form a cohesive structure and enable comparison on a common scale. Table 5 shows descriptive statistics for the fair scores for 2009 and 2010 which, as expected, correlated highly with raw scores, 0.957 and 0.856, respectively.

The potential impact of ‘fair scores’ on candidate selection could be compared with the raw scores. Traditionally Dundee medical school selects the top 330. If the fair scores were used instead of the raw scores, the outcome of 28 (6.2%) and 46 (9.6%) candidates would have been changed in 2009 and 2010, respectively. All candidates whose selection would have been impacted by the fair scores were within 1 SD of the cut-off point of the raw scores (Table 7).

Discussion

This article has sought to report on a complex process using a number of sources of data as well as qualitative information. As a case study in one medical school, it is highly context specific and our findings may not generalise well to other settings. For instance, staff and student attitudes may differ regionally and the psychometric performance of an MMI is likely to be contingent upon the approach used for station development and testing. However, there is no reason why the candidates interviewed through this process should not be broadly representative of all UK applicants.

Table 7. Descriptive statistics for MFRM fair scores.

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>105.8 (12.1)</td>
<td>106.7 (12.7)</td>
</tr>
<tr>
<td>Range</td>
<td>61.9–141.9</td>
<td>54.0–136.5</td>
</tr>
<tr>
<td>Mean difference</td>
<td>–0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Range of differences</td>
<td>–11.9–7.0</td>
<td>–10.6–8.8</td>
</tr>
<tr>
<td>Selection difference</td>
<td>28 (6.2%)</td>
<td>46 (9.6%)</td>
</tr>
</tbody>
</table>

Progression from a traditional interview to a full-scale MMI

The strategy of generating institutional support through staged introduction, presentation of local and international evidences to the admissions committee and at open meetings, and including organisational leaders at major decision points proved effective. The MMI at Dundee has demonstrated generally sound psychometric properties, similar to the results published by others (Eva et al. 2004; Lemay et al. 2007; Reiter et al. 2007). Hence, there is now apparent acceptance of the process by applicants in general, by interviewees in particular and by assessors. Recruitment of assessors was a daunting prospect initially, but did not prove as difficult as expected. The use of students and simulated patients eased the burden on staff and appears valuable to the selection process, as each group conveys a different, yet relevant, perspective on candidates’ qualities. The interactive, student led stations performed at least as effectively as faculty run interviews and have strong face validity, as confirmed by follow-up questionnaire. Though concerns have been raised regarding disclosure of station content by students our experience does not support this (Nicholson & Jaiyesimi 2010). Reassuringly, Reiter et al. (2006) have reported no significant differences in mean scores for candidates who were interviewed after receiving MMI station stems. While the lack of drift in mean scores across the MMI days and years also suggests that breaches of confidentiality did not translate into an overall improved performance, advantage to some candidates cannot be excluded. It is a feature of competitive selection that training and coaching will emerge but the impact of this on a performance test is likely to be less than that on traditional interviews.

We recognise that the construct validity of our MMI is largely unknown because there are few separate measures for most of the attributes measured in medical school (except communication). To some extent, this is a separate issue; i.e. the blueprint for what medical schools teach and test may lag behind or be imperfectly aligned with explicitly defined institutional values to which the MMI station content was mapped. For example, most medical schools purport to value empathy and integrity, yet few assess these skills rigorously on
an individual basis. Once the first cohort of students who performed in the MMI have reached their 4th or 5th year, if overall MMI scores do predict a significant portion of the variance in medical school performance, the issue of construct validity will become less important. One benefit to emerge from our ability to assess candidates according to individual domains, albeit with limited reliability, has been its value as a simple form of specific feedback for rejected applicants. On a pragmatic basis, this has proved very useful.

Reliability
Overall reliability for both years was adequate and consistent. The removal of the weakest station would have improved reliability, suggesting that detailed, question specific analysis and feedback are required on all new stations to identify and rectify problems, ideally within cycle. Reliability by station type was comparable, with both student-assessed and staff-assessed stations performing well. However, students perform better than staff in terms of being less dovish and making more use of the rating scales full range.

Differences in performance between applicant groups are not the focus of this article, but we have included some data to illustrate the types of comparisons that become possible once a reliable process is in place. That graduates performed consistently better than others is not surprising as age is associated with superior performance in OSCEs (Calvert et al. 2009), they have increased motivation relative to school leavers (Wilkinson et al. 2004) and they perform better on the course initially (Price & Wright 2010). While the evidence for the association with age and performance in medical school is not conclusive within the UK medical education context, it is perhaps evidence of concurrent validity that they perform better in an MMI.

While further research must be conducted in order to pinpoint the underlying causes and effects, it is also important to recognise that other measures in selection, (such as the UKCAT where age has been shown to be negatively correlated with performance) may have contrary effects (UKCAT Annual Report 2008). Using a combination of methods may result in a more ‘level playing field’ for all candidates. However, it should also be noted that while the mean score difference reached statistical significance, the overall numerical impact was small as the proportion of graduate/mature candidates was low. Alternatively, if MMIs can identify those who will consistently perform better in medical school and beyond in the UK as elsewhere, it is surely appropriate that it is used in preference to other less valid admissions tools, irrespective of its impact on the demography of entrants. Thus, it will be important to establish if maturity continues to have the same impact on performance at medical school exit and in postgraduate years.

Views of applicants and assessors
Data from the follow-up questionnaire proved reassuring and useful, primarily for convincing staff that the task was not too challenging for applicants. It also guided station development by identifying specific scoring challenges and leading us to limit station content to medically relevant topics. Together with station-specific reliability data, it also suggests that more specific training may benefit certain interactive stations. However, striking an effective balance between training needs and assessors’ willingness to devote time to training is a perennial challenge for the interviewing process and not unique to the MMI.

Further research
MFRM shows that adjusting for assessor leniency/stringency as source of error has a noticeable impact on selection and merits consideration. And while the statistics show that the MMI generally performed well in both years, they do raise a number of important issues that warrant further investigation. For instance, the need for variation in station content or merit of assessing potential bias using measures of differential item functioning.

With the weight of existing evidence that the MMI is an improvement on traditional methods of interviewing, and that it can also be implemented successfully within the space of a few years, this method is likely to be picked up by other schools. Evaluation of the MMI’s predictive validity and further assessment of its psychometric properties would be enhanced through collaborative research with other institutions. For example, we could begin tracking larger numbers of students so subgroup effects such as age could be determined with greater accuracy and look at performance of candidates who did not receive an offer from one or more schools. We also would like to explore the feasibility of developing a shared bank of MMI station content, as has recently been done successfully elsewhere (Roberts et al. 2009). Given the significant challenge of developing and piloting reliable stations, this would allow institutions to benefit from a larger sample of stations with performance data and minimise the impact of content leakage and candidate ‘grooming’. To do so, a common framework will need to be established (duration, domains, scoring scale).

Despite the considerable intellectual and logistical challenge, MMIs were successfully introduced at the University of Dundee. Our interviews were well accepted by staff and candidates alike and demonstrated robust psychometric properties. This was achieved at an acceptable standard within 2 years. There is significant scope to improve and understand the measurement issues further.

Acknowledgements
The authors are indebted to Harold Reiter, Kevin Eva and Geoffrey Norman for allowing the first author to visit and observe the MMI at McMaster University; to Martin Pippard and Gary Mires for faculty support in transitioning to the MMI; Gominda Ponnampenula for allowing us to adapt his work for use in two of our stations; to Lynda Cochrane and Ben Kumwenda for assistance with data collection and analysis; and to Gordon Black for working intensively in recruitment of assessors and tracking of candidates and assessors throughout the process.

303
Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Notes on contributors

JON DOWELL, BM, BS MRCGP, MD, is Admissions Convener and Head of Tayside Centre for General Practice.

BONNIE LYNCH, PhD, is lecturer in Medical Education.

HETTIE TILL, PhD, is Lecturer in Assessment (Psychometrician), Centre for Medical Education.

BEN KUMWENDA, MSc, is Research Administrator in Medical Education.

ADRIAN HUSBANDS, MSc, is lecturer in Medical Education.

References


UK Clinical Aptitude Test. UKCAT home. Available from: http://www.ukcat.ac.uk/home/
