Reliability estimates: behavioural stations and questionnaires in medical school admissions

Naomi Gafni,[1] Avital Moshinsky,[1] Orit Eisenberg,[1] David Zeigler[1] & Amitai Ziv[2]

CONTEXT Assessment centres used in evaluating the non-cognitive attributes of medical school candidates must generate scores that reflect as accurate a measurement as possible of these attributes. Thus far, reliability coefficients for such centres have been based on limited samples and individual administrations, without reference to the error of variance that may result from retesting, or from the existence of multiple centres designed to measure the same attributes.

METHODS The National Institute for Testing and Evaluation in Israel has developed and administered two assessment centres: MOR is used by two medical schools and one dental school, and MIRKAM by another medical school. Each centre comprises eight or nine behavioural stations, a standardised biographical questionnaire, and a judgement and decision-making questionnaire. We calculated generalisability coefficients for each centre's eight or nine stations by year, composite reliability coefficients for the overall assessment centres, test–retest correlation coefficients for repeaters, and a correlation coefficient between the centres.

RESULTS Between 2006 and 2009, 2662 and 2023 examinees participated in MOR and MIRKAM, respectively; 1479 of these participated in both. The average generalisability coefficients for the stations were 0.69 for MOR and 0.67 for MIRKAM. The composite reliability coefficients for the full centres (behavioural stations plus questionnaires) were 0.79 and 0.76 for MOR and MIRKAM, respectively. The correlations for repeaters, corrected for restriction of range, were 0.59 and 0.43 for MOR and MIRKAM stations, respectively, and 0.72 and 0.65 for the full MOR and MIRKAM assessments, respectively. The correlation between scores on the MOR and MIRKAM stations was 0.56 (0.75 for the overall score).

DISCUSSION The minimal reliability desirable for high-stakes decision making (0.80) was obtained only for 14 or 15 stations with questionnaires. Nevertheless, the values obtained are considerably higher than reliability coefficients for single interviews. The questionnaires contribute significantly to the accuracy of the measurement. These reliability measures constitute an upper threshold for measures of validity.

Medical Education 2012: ***46****: 277–288*
doi:10.1111/j.1365-2923.2011.04155.x

[1]National Institute for Testing and Evaluation (NITE), Jerusalem, Israel
[2]Israel Centre for Medical Simulation, Sheba Medical Center, Tel Hashomer, Israel

Correspondence: Naomi Gafni, National Institute for Testing and Evaluation R&D, PO Box 26015, Jerusalem 91260, Israel. Tel: 00 972 2 675 9506; Fax: 00 972 2 675 9543; E-mail: naomi@nite.org.il

INTRODUCTION

Until recently, most medical schools relied on an individual interview as the sole means of evaluating a candidate's non-cognitive attributes.[1–3] In the last 6 years, however, alternative admissions systems based on multiple, structured, evaluations of non-cognitive parameters have been introduced.[4–7] McMaster University in Canada was the first to eschew the traditional interview in favour of the multiple mini-interview (MMI),[5] a measurement tool styled after the objective structured clinical examination (OSCE). Research suggests that these new techniques yield summary scores that have reasonable reliability (0.6–0.81).[5,7–11]

A satisfactory level of reliability is particularly important in high-stakes admissions systems because it informs us of an upper limit of validity that can be achieved. Although several validity studies of the MMI technique have been conducted,[12–14] these have been based on fairly limited data obtained from small samples. Thus far, the research has presented G-coefficients obtained for candidates participating in eight to twelve MMI stations in one administration. These precision estimates have various limitations:

1 they do not include error variance attributable to occasion;
2 they represent systems with specific numbers of MMI stations and estimates based on D studies, but do not provide estimates based on actual larger numbers of stations, and[7,15,16]
3 they do not include error variance attributable to design variance.

The MMI design is defined as the particular composition of stations comprising an MMI system (i.e. a specific mix of group stations versus individual stations; stations using standardised patients [SPs] versus stations not using SPs; stations using simulations versus stations using conventional interviews, etc.). What remains to be determined is the degree to which a score derived from a specific MMI design is generalisable to other MMIs developed by different teams (e.g. an MMI designed by and for another university admissions system that supposedly measures the same personal attributes).

The objective of this study is to provide some MMI reliability estimates that include the sources of variance mentioned above: occasion and design (but not the exact variance component related to each one of these sources), as well as G-coefficients for 16 and 17 behavioural stations. Each of these estimates is an approximation of the degree to which replication of the measurement procedure, within the same universe of generalisation, will yield the same score.

The National Institute for Testing and Evaluation (NITE), Jerusalem, is responsible for the development, administration and scoring of two admissions systems used by three medical schools and one dental school. In 2003, inspired by the MMI model, Tel Aviv University's Sackler School of Medicine, along with NITE and the Israel Centre for Medical Simulation (MSR), developed an assessment centre named MOR, which utilises a diverse array of evaluation tools,[3] including eight MMI-like stations (nine, as of 2008) and two customised questionnaires (the Judgement and Decision-making Questionnaire [JDQ] and the objective standardised Biographical Questionnaire [BQ]), that focus on candidates' judgement and decision-making abilities and on their background. The new admissions system was first implemented in 2004. The Technion School of Medicine in Haifa and the School of Dental Medicine at the Hebrew University in Jerusalem joined the project in 2006 and 2007, respectively. In 2006, the School of Medicine at the Hebrew University in Jerusalem, in collaboration with NITE, developed its own assessment centre. Known as MIRKAM, this assessment centre utilises eight mini-interviews, as well as the two questionnaires used in MOR. The MOR and MIRKAM systems use the same questionnaires and have similar evaluation forms for the stations, but the nature of the behavioural stations is different in each.

The development and implementation of the assessment centres have yielded data for the following groups of applicants:

1 applicants who participated in MOR;
2 applicants who participated in MIRKAM;
3 applicants who participated in both MOR and MIRKAM;
4 repeaters of MOR;
5 repeaters of MIRKAM, and
6 repeaters of both MOR and MIRKAM.

The objective of this study is to estimate the G-coefficients within each centre (with eight or nine stations per year), across both centres (with 16 or 17 stations), and of the stations (either eight or nine, or 16 or 17) combined with the two questionnaires. In addition, the data also make it possible to estimate the test–retest reliability of all the above entities. This estimate takes into consideration the variance

attributable to occasion. Such an estimate determines the degree to which scores on a particular MMI (containing eight or nine, or 16 or 17 stations) can predict relevant future performance. In a certain sense, test–retest events seem to constitute better replications drawn from the universe of generalisation than 'replications' based on a single administration. These test–retest replications, however, are not perfect because those who retake the test are not representative of the whole population of examinees. In conclusion, we are dealing with approximations, each of which suffers from several limitations that must be acknowledged.

An additional estimate of consistency is the correlation between the two sets of MMIs (MOR and MIRKAM) taken by the same students in a certain year. Such an estimate takes into consideration mainly the variance attributable to the MMI design and, to some degree, variance attributable to occasion (the two systems are run on different days, albeit within 3 months of one another). It also indicates the degree to which we can expect a particular MMI score to predict future performance in terms of the relevant attributes that are supposedly measured by that MMI.

These analyses provide evidence regarding the various reliability estimates derived from each dataset, and help to determine whether the reliability of an admissions system that includes a certain number of mini-interviews can be increased, either by adding more stations or by supplementing them with additional measures. Given that, in addition to the MMI, both MOR and MIRKAM utilise two questionnaires, it would be interesting to examine the reliability of the full assessment centre (MMI + questionnaires) for all of the above datasets and compare the results with the reliability estimates for the MMI only.

METHODS

Both the MOR and MIRKAM assessment centres consist of three assessment tools: the JDQ; the BQ, and a set of behavioural stations (which differ between MOR and MIRKAM). During 2004–2007, MOR included eight behavioural stations; in 2008 one station was added, bringing the total to nine. Since its inception in 2006, MIRKAM has included eight interview stations. Since 2006, candidates who participated in both assessment centres have completed the questionnaires only once a year and their scores have been used for both systems. MOR and its components have been described in detail by Ziv *et al.*[4]

Behavioural profile

A survey of medical professionals, as well as a review of bulletins directed at medical school candidates, yielded the following list of attributes required of a 'good doctor': an ethical approach; integrity; professional responsibility; empathy; a sense of service; commitment to the patient; interpersonal communication skills; self-confidence; human sensitivity; the ability to discern details; the ability to identify the need for help and to seek and accept help; openness; initiative; a positive attitude towards authority; self-awareness; maturity, and the ability to cope with pressure.

The attributes deemed feasible for evaluation were as follows: (i) interpersonal communication ability; (ii) ability to handle stress; (iii) initiative and responsibility, and (iv) self-awareness and maturity. Both steering committees adopted these four dimensions. The content of the behavioural stations, however, differed somewhat between MOR and MIRKAM, as will be elaborated below.

The Judgement and Decision-making Questionnaire

The JDQ involved three essay-writing tasks describing real-life ethical dilemmas (for which a total of 45 minutes was allocated). Each task required respondents to analyse a complex situation and to then reach, and justify, a decision.

The scoring system was designed to yield a score that reflected the candidate's perceptions of moral complexity. For each task, the arguments provided by the candidate were counted; each argument received 0–2 points based on how it related to general values, and on professional and moral considerations. The essays were assessed by two independent raters, and the final score represented the average of the two evaluations. In cases of a substantial difference between ratings, a third expert's assessment was weighted in the final score. The raters were professionals in the social sciences who were thoroughly trained to score the questionnaires. The range of G-coefficients for the JDQ across the four cohorts was 0.64–0.68.

The Biographical Questionnaire

The BQ consisted of 20 open-ended questions arranged in two sections. The first section focused on the candidate's life experience, including activities during high school and afterwards (military service, volunteer work, etc.). The second section focused on

the candidate's emotional awareness and included items pertaining to past experience in coping with challenging emotional situations. The total time allocated was 95–120 minutes (depending on the year of administration).

Scoring was based on factors indicating honesty, credibility, responsibility, persistence, self-evaluation and interpersonal behaviour. Each question was scored on a scale of 1–5, according to a predefined detailed scoring rubric of accepted responses. Candidates' responses were assessed by two independent raters, as in the JDQ. The same group of raters scored both questionnaires. The range of G-coefficients for the BQ across the four cohorts was 0.64–0.71.

MOR: simulation stations

The MOR included eight (until 2007) or nine (from 2008) individual stations (6–9 minutes each) and two group stations (30 minutes each).

Three simulation stations each consisted of a challenging encounter between a candidate and an SP in which the candidate played the role of an interlocutor in a medical or non-medical context that did not require prior medical expertise. For example, one situation involved an angry 'patient' whom the candidate was required to calm down. In 2009, one simulation station was replaced by an interview station in which an ethical issue was discussed.

Two debriefing stations required the candidate to be interviewed by a rater who had observed his or her behaviour in the simulation stations and now asked structured questions relating to the candidate's performance. In 2008, one debriefing station was replaced by an interview station.

One standardised personal interview station (two from 2009) involved a short personal interview focused on the candidate's attitudes towards the medical profession and on current issues in medical policy.

In each of the two group stations, a group of six candidates was required to perform a task conjointly (e.g. rating the importance of statements related to the medical profession). In 2008, one group station was replaced by a team station which involved an encounter between a team of three candidates and an actor playing a role related to a relevant situation. The candidates faced interpersonal and intra-group challenges.

Candidates' attributes were scored using a structured assessment form that focused on personal attributes within the four domains cited earlier. Each quality was represented by between two and five different items, each rated on a scale of 1–6. In two (or three) simulation stations, scoring was performed by both a faculty member and an SP. In the remaining stations, scoring was performed by a single faculty member (for reasons of cost and other logistical considerations). The raters and SPs changed from station to station, except for the pairs of simulation and debriefing stations, where the faculty member who had observed and rated the candidate in the simulation station also conducted the debriefing. Raters and SPs changed from one administration to another. The assignment of faculty members and SPs to the different candidates was random. A total of 348–386 faculty members and 33–35 SPs participate in MOR every year. The final score was the average of the scores generated for each personal quality across stations.

MIRKAM: multiple mini-interviews

Unlike in MOR, the eight stations in MIRKAM all consisted of individual structured interviews. Each station lasted 10 minutes.

In three semi-simulation stations, the interviewer and the candidate engaged in role-play related to a conflict between two characters (e.g. a doctor and a patient). After 5 minutes, the interviewer stopped the role-play and commenced the debriefing process. The debriefing was based on questions similar to those used in MOR.

In two stations, the candidate was presented with an ethical medical dilemma to discuss.

In three structured personal interview stations, the candidate was asked about his or her biographical history.

In each station, candidates were scored by a single interviewer (a faculty member [a total of eight independent interviewers across stations]) using an assessment form similar to that used in MOR. This form consisted of one content component (the extent to which the answers corresponded to a specified 'expected best answer') and three personal attribute components (also used in MOR), which covered, respectively: interpersonal communication skills; the ability to handle stress, and maturity and self-awareness. Interviewers were randomly assigned to stations in each administration. Between 140 and

180 faculty members participate in MIRKAM each year.

The stations for MOR and MIRKAM, as well as the questionnaire items, were developed by professional test developers from NITE and were reviewed and approved by members of the admissions committees of each institution.

Both MOR and MIRKAM are believed to measure dimensions relevant to the future professional behaviour of doctors. Despite the variations between the two assessment centres, they are designed to measure the same dimensions, albeit by somewhat different operational means. As such, the degree of correlation between MOR and MIRKAM is expected to be stronger than the correlation between either of them and any future criterion.

Scoring

In each assessment centre, four scores were calculated for each candidate: (i) a JDQ score; (ii) a BQ score; (iii) a behavioural stations score, derived from the weighted average of scores on the four domains of attributes mentioned above, and (iv) a total score derived from an average of the preceding three scores weighted 20/20/60, respectively, for MOR, and an average of the preceding three scores weighted 15/25/60, respectively, for MIRKAM.

The relative weight of each component was assigned by the steering committee. The highest weight was assigned to the hands-on behavioural component, which accounted for 50% of the total time spent in the assessment centre and 11 of the total of 15 or 16 independent evaluations for MOR, and eight of the 12 independent evaluations for MIRKAM. For both MOR and MIRKAM, the raw scores were converted to standard scores, using a scale with a mean of 200 and a standard deviation of 20 (minimum: 150; maximum: 250).

Setting

Since 2006, candidates applying to 6-year programmes in at least one of four institutions (three medical schools and one dental school) have been invited to participate in an assessment centre, provided that their average matriculation score and Psychometric Entrance Test (PET) score were above the cut-off determined by the admissions committee of the relevant school.[17] Candidates for three of the schools participated in MOR and candidates for one school participated in MIRKAM. The behavioural stations of MOR were administered at the Israel Centre for Medical Simulation, and those of MIRKAM were administered at a different appropriate facility each year. The administration of MOR was conducted over 4–6 days per year. The administration of MIRKAM was conducted over 10–20 days each year. The two questionnaires were completed by candidates on a separate occasion at one of the university campuses.

Candidates completed the questionnaires only once each year, but participated in either or both of the assessment centres, depending on the schools to which they had applied. Strict ethical principles were adhered to throughout the development and implementation of both MOR and MIRKAM. Strict anonymity of data was maintained in order to protect the participants' privacy. All results were reported in an aggregated manner.[18]

Participants

Table 1 presents the number of participants in each of the two assessment centres, as well as the number of candidates who participated in both assessment centres each year (2006–2009) and across the 4 years. The mean age of the candidates was 22 years and the male : female ratio of the entire cohort was 48 : 52.

Reliability estimation

In order to compare the reliability estimates of MOR and MIRKAM with previous findings published in the literature, it was decided to score each station for each candidate and then estimate the generalisability of the MMI based on these eight or nine station scores. The scoring rubric for each station consisted of six to nine criteria scored on a 6-point scale (1, 2 = low; 3, 4 = average; 5, 6 = high). The station score was defined as the sum of the scores across these criteria. It should be noted that the intercorrelations among the various criteria were high, thus justifying the use of the sum of the scores as the station score

Table 1 Number of participants in each assessment centre, by year and across years

Assessment centre	2006	2007	2008	2009	Total
MOR	565	645	695	757	2662
MIRKAM	413	451	533	626	2023
MOR + MIRKAM	308	341	415	415	1479

(e.g. in 2008, the inter-item reliabilities of the nine stations were 0.88–0.94).

Different test versions were used on each evaluation day. A partial list of the sources of error variance includes the following: scenario/question; interviewer; SP; composition of the group or team in a group or team station; order of stations; number of interviewers in each station, and all interactions among candidates, interviewers and the other variables. As all of these variables were confounded, it was impossible to model the relative contribution of each to the error variance. Therefore, in order to determine the reliability of stations as a whole, a candidate × station analysis of variance (ANOVA) was performed for each year.

Table 2 Summary of effects, estimated variance components and the G-coefficient for MOR 2006

Effect	d.f.	MS	Estimated variance
Candidate	564	191.29	17.337
Station	7	46 236.45	81.743
Candidate*station	3948	52.593	52.593
G-coefficient			
$\sigma^2_{(candidate)}/(\sigma^2_{(candidate)} + (\sigma^2_{(candidate*station)}/8)) = 0.73$			

MS = mean square

RESULTS

Generalisability of the behavioural stations

Table 2 reports the components of variance and illustrates an overall test generalisability (i.e. the reliability of the average of all ratings) of 0.73 for MOR behavioural stations in 2006 (this is similar to that presented by Eva *et al.*[5]). This coefficient is an analogue of a reliability coefficient in classical theory, which tells us the extent to which an individual's position within a score distribution remains stable across items (as opposed to a coefficient based on absolute error variance which is used when a minimal qualifying score is predetermined).

Table 3 presents the average G-estimates for the behavioural stations of MOR and MIRKAM for each year and across years, as well as the composite reliability coefficients for the full assessment centres.

As two pairs of stations included in MOR were rated by the same rater (the simulation with subsequent debriefing stations), an additional estimate was computed for six or seven stations (instead of eight or nine stations), treating the pair of stations as one. The second row in Table 3 presents these estimates. As the two stations do not assess exactly the same dimension, we believe that the 'true' G-coefficient of the MOR behavioural stations falls somewhere between these two estimates.

Table 3 G-coefficients of the behavioural stations and composite reliability coefficients for the full assessment centres of MOR and MIRKAM, by year and across versions

	G-coefficients				
Assessment centre	**2006**	**2007**	**2008**	**2009**	**Mean across years**
MOR (eight stations in 2006/2007; nine in 2008/2009)	0.73	0.76	0.79	0.77	0.76
MOR (six stations in 2006/2007; seven in 2008/2009)	0.64	0.71	0.72	0.69	0.69
MIRKAM (eight stations)	0.74	0.68	0.63	0.61	0.67
Full assessment centre (stations + questionnaires)	**Composite reliability coefficients**				
MOR (eight stations in 2006/2007; nine in 2008/2009)	0.83	0.83	0.84	0.82	0.83
MOR (six stations in 2006/2007; seven in 2008/2009)	0.78	0.80	0.80	0.78	0.79
MIRKAM (eight stations)	0.83	0.76	0.75	0.70	0.76

The G-coefficients are similar to the reported range for MMI of 0.65–0.81,[4,7,11] although for a considerably larger sample (over 4000). In general, the estimates are somewhat higher for the MOR behavioural stations than for the MIRKAM behavioural stations. (The consistent decline in the G-coefficients for MIRKAM may be attributable to some degree to a lower participation rate in MIRKAM rater-training sessions.)

Reliability of the full assessment centre

Both MOR and MIRKAM comprise stations and two questionnaires. The reliability estimate of the full assessment centre was obtained by calculating the reliability of a composite score.[19–21] Table 3 presents reliability coefficients for each assessment centre in its entirety. The total assessment centre score is an average of the two questionnaire scores (the BQ and the JDQ) and the behavioural stations score, weighted at 20/20/60 and 15/25/60 for MOR and MIRKAM, respectively. The two questionnaires increased the reliability of MOR and MIRKAM considerably. These estimates are considered reasonable for the measurement of non-cognitive attributes.

Generalisability of all MOR and MIRKAM behavioural stations combined

According to Roberts *et al.*,[8] who conducted a D study on expected changes in reliability when the number of MMI questions and associated testing time were increased, in order to obtain a reliability of 0.8, it is necessary to administer either 14 stations with one interviewer per case, or 12 stations with two interviewers per case. The data available from candidates who participated in both MOR and MIRKAM allow for the computation of a G-coefficient for the combined MMI of the two systems (14–17 MMIs). Although the two MMI systems are not identical, they are intended to measure the same personal attributes using similar rating forms, thereby justifying the use of such a procedure. A total of 1479 candidates participated in both MOR and MIRKAM (Table 1).

Table 4 presents G-coefficients for the combined behavioural stations of MOR and MIRKAM, based on those candidates who participated in both systems in the same year. The first row presents coefficients for 16 (2006 and 2007) and 17 (2008 and 2009) stations, disregarding the dependency between two pairs of stations included in MOR. The second row presents the same information with each dependent pair of stations treated as one station.

The increase in the number of stations in 2008 was not accompanied by an increase in the generalisability coefficient. This might be attributable to changes in one or all of the following: MOR scenarios; SPs; raters, and examinee population.

As expected, increasing the number of stations to 14 was found to be associated with an increase in reliability from 0.67 (MIRKAM) or 0.76/0.69 (MOR [Table 3]) to 0.81/0.77 across the years. It seems that the most probable generalisability estimate of the combined behavioural stations score across the years would be the average of the two coefficients (0.81 and 0.77): 0.79. This estimate parallels that of Roberts *et al.*[8] for 13 stations. Roberts *et al.*,[8] however, did not include any variance attributable to occasion, whereas in the current study, some variance is attributable to occasion, as well as to other factors that vary from one MMI system to the other.

Composite reliability of all MOR and MIRKAM behavioural stations combined

The last two rows of Table 4 present the reliability coefficients of the combined MOR and MIRKAM systems, including the behavioural stations and the

Table 4 G-coefficients for the combined multiple mini-interview stations of MOR and MIRKAM

G-coefficients for MOR and MIRKAM combined	2006	2007	2008	2009	Mean
Behavioural stations					
MOR + MIRKAM (16 or 17 stations)	0.80	0.81	0.81	0.80	0.81
MOR + MIRKAM (14 or 15 stations)	0.76	0.79	0.77	0.76	0.77
Full assessment centre (stations + questionnaires)					
16 or 17 stations + questionnaires	0.84	0.85	0.84	0.84	0.84
14 or 15 stations + questionnaires	0.82	0.84	0.82	0.82	0.83

two questionnaires (used in both systems). It can be concluded that increasing the number of stations to 14 is associated with an increase in reliability from 0.79 (Table 3) to 0.83 for the entire MOR system, and from 0.76 (Table 3) to 0.83 for the entire MIRKAM system, across the years.

Test–retest reliability for the behavioural stations score and for the full assessment centre

Applicants who are not admitted to medical studies may choose to reapply in the following year and may be invited to repeat MOR and/or MIRKAM. Thus far, 405 applicants have repeated MOR and 230 applicants have repeated MIRKAM. MOR became operational in 2004 and 71 repeaters had participated in MOR for the first time prior to 2007. Candidates retaking the test may have learned from the experience of the first test; they know what to expect and may be better prepared for the retest. Therefore, retest scores are better than those on the original test (by 0.5 of a SD). The degree of improvement is not related to the stability of score ranking, to which the test–retest correlation attests. Table 5 presents observed test–retest correlation coefficients for the MMI stations and for the full assessment centre for these repeaters. However, these correlations constitute an underestimation of the 'true' test–retest correlations because they are based on a restricted population of only those applicants who failed to achieve the cut-off admissions score. Table 5 also presents the corrected (for range restriction) correlations.[22] The correction provides an estimated correlation were all the candidates to retake the test.

Both the G-coefficient and the test–retest coefficient indicate, in terms of variance ratios, the degree to which replications of the measurement procedures are related to one another. In the second case, the universe of generalisation includes an additional source of variance: occasion. As expected, the test–retest reliability estimates are lower than the inter-station reliabilities, which are based on only one administration (0.59 versus 0.76/0.69 for MOR stations; 0.43 versus 0.67 for MIRKAM stations; 0.72 versus 0.83/0.79 for the full MOR assessment centre; 0.65 versus 0.79 for the full MIRKAM assessment centre). Thus, the closest approximation of the 'true' reliability of eight MMI stations seems to be about 0.10–0.25 lower than the inter-station G-estimate, and the best reliability estimate for a full assessment centre is about 0.09–0.15 lower than its single administration estimate. It should be noted that the test–retest correlations presented above were obtained for candidates who were retested within a year; these correlations are expected to become lower as the time interval increases.

Test–retest reliability for the combined behavioural stations score and for the full assessment centre, based on 16 or 17 stations

A total of 135 candidates participated twice in both MOR and MIRKAM during the years 2006–2009. Table 6 presents the following test–retest correlations for this group for the combined 16 or 17 behavioural stations of MOR and MIRKAM, for the combined total score of MOR and MIRKAM (16 or 17 stations + questionnaires), for the MOR behavioural stations, for the MOR total score, for the MIRKAM behavioural stations, and for the MIRKAM total score. The last four correlations are provided for the sake of comparability within the same group.

The estimates presented in the table indicate that increasing the number of stations from eight to 16 or 17 raises the test–retest reliability of the behavioural stations from 0.59 (MOR) and 0.48 (MIRKAM) to 0.71, and that of the entire assessment centre from 0.68 (MOR) and 0.70 (MIRKAM) to 0.80. It seems

Table 5 Observed and corrected (in parentheses) test–retest correlations for repeaters of MOR and MIRKAM behavioural stations, and for the full assessment centre (including the questionnaires)

	2007		2008		2009		Across years	
	n	Correlation	*n*	Correlation	*n*	Correlation	*n*	Correlation
MOR stations	100	0.53 (0.62)	112	0.55 (0.64)	122	0.43 (0.51)	405	0.50 (0.59)
MIRKAM stations	52	0.43 (0.54)	70	0.36 (0.39)	108	0.28 (0.38)	230	0.34 (0.43)
MOR total score	100	0.64 (0.72)	112	0.67 (0.75)	122	0.51 (0.67)	405	0.61 (0.72)
MIRKAM total score	52	0.64 (0.74)	70	0.52 (0.56)	106	0.47 (0.61)	230	0.52 (0.65)

Table 6 Observed and corrected test–retest correlations for 135 candidates who participated in both MOR and MIRKAM twice

	Observed correlation	Corrected correlation
Combined stations	0.58	0.71
Combined total	0.65	0.80
MOR stations	0.48	0.59
MOR total	0.54	0.68
MIRKAM stations	0.38	0.48
MIRKAM total	0.57	0.70

that eight stations alone do not equal the precision of other measures commonly used to facilitate admissions decisions.

The correlation between MOR and MIRKAM

In the years 2006–2009, 1479 candidates participated in both MOR and MIRKAM. As mentioned above, the two systems differed only in their behavioural stations. Table 7 presents the intercorrelations among the various variables within and across the two systems.

The pattern of the correlations in Table 7 supports expectations, given that the two assessment centres supposedly measure similar constructs. That is, the correlations of the two questionnaires with the two sets of stations are similar (0.28 and 0.23 between JDQ and the behavioural stations for MOR and MIRKAM, respectively; 0.51 and 0.47 between BQ and the behavioural stations for MOR and MIRKAM, respectively). Likewise, the correlation between MIRKAM stations and MOR total score (0.59) is similar to the correlation between MOR stations and MIRKAM total score (0.62). The correlations between the various components and the total scores reflect the respective weights of the components in the total score.

The Pearson correlation coefficient between the MOR and MIRKAM behavioural station scores across the 4 years was 0.56. Assuming that MOR and MIRKAM are more similar to one another than either of them is to any criterion variable used in any validity study, the correlation between them sets an upper limit for any validity correlation coefficient that may be found. The correlation between the two total scores was 0.75; however, given that both were based on scores on the same two questionnaires, a high correlation is to be expected.

DISCUSSION

Thus far, the research concerning the use of the MMI as a selection tool for medical studies has focused on generalisability estimates based on one administration only. As there is fairly limited evidence of the predictive ability of this instrument, it is particularly important to investigate the degree to which one MMI score can predict another MMI score obtained in a different administration. The data gathered at NITE make it possible to calculate previously unobtainable reliability estimates for the MMI component of medical school admissions. Using large samples of participants over 4 years in two assessment centres that contained two variations of the MMI system, we confirmed previous reliability estimates based on generalisability theory. The G-coefficient estimates found in this study varied somewhat from year to year and from one MMI system to another. The estimates also varied according to the method of estimation. Thus, the reliability of the MOR behavioural stations was somewhat lower when dependency between

Table 7 Observed correlations between MOR and MIRKAM stations, MOR and MIRKAM total scores, and questionnaire scores

Variable	JDQ	BQ	MOR stations	MIRKAM stations	MOR Total score
BQ	0.35				
MOR stations	0.28	0.51			
MIRKAM stations	0.23	0.47	0.56		
MOR total score	0.53	0.72	0.93	0.59	
MIRKAM total score	0.46	0.72	0.62	0.93	0.75

JDQ = Judgement and Decision-making Questionnaire; BQ = Biographical Questionnaire

stations was accounted for. The best estimate of single administration reliability for eight or nine behavioural stations seems to be the average G-coefficient across 4 years, 4685 participants, and across the two assessment centres: namely, 0.69.

Both MMI systems, the MOR and MIRKAM, include two questionnaires, a judgement and decision-making questionnaire and an objective standardised biographical questionnaire, in addition to the behavioural stations. This paper has focused on the reliability of the behavioural stations and the total score of the assessment centre, rather than on the questionnaires. However, it can be concluded from the results that including the questionnaires was associated with an increase of about 0.1 in the reliability coefficient. According to Roberts *et al.*,[8] this increase in reliability parallels the increase in reliability associated with increasing the number of behavioural stations from eight to 14 (0.70–0.80).

Combining the behavioural stations of the two systems allows us to compute the G-coefficients of 14 or 15 behavioural stations. These estimates varied from 0.76 to 0.79 (accounting for dependency between stations). Thus, increasing the number of the behavioural stations from six or seven to 14 or 15 was associated with an increase of 0.02–0.15 (depending on the year). Combining the behavioural stations also increased the reliability of the total score (including the questionnaires) from 0.79 (MOR) and 0.76 (MIRKAM) to 0.83, reflecting increases of 0.04 and 0.07 for MOR and MIRKAM, respectively. As expected, augmenting the admissions method by including the questionnaires is more meaningful when the number of stations is relatively small.

Doubling the number of stations would increase the level of reliability to meet the standards required for selection purposes. This, however, does not appear to be practical as a result of the concurrent increase in evaluation time and cost. Asking candidates to complete appropriate questionnaires raises the reliability to an acceptable level and constitutes a good means of improving the psychometric quality of the admissions procedure.

The reliability coefficients estimated for a specific administration do not contain error variance related to occasion or other factors that change from one year to another and, therefore, tend to overestimate the true reliability of the selection method. The test–retest correlation coefficients computed for candidates who repeated the assessment centre (as well as the correlations between two variations of the MMI method) better represent the true reliability of the admissions system.[23,24] They set an upper limit for validity coefficients, which are probably based on criterion measures that differ from the admissions procedure far more than two variations of the same procedure. However, these measures are computed based only on a sample of candidates who were not admitted and therefore underestimate the true reliability coefficients. After correcting for curtailment, the resulting coefficients for the behavioural stations were about 0.10 (for MOR) to 0.24 (for MIRKAM) lower than the G-estimates. The corrected test–retest coefficients for the total score were about 0.07 (for MOR) and 0.11 (for MIRKAM) lower than the single-administration reliability coefficients.

Obviously, the most accurate measure is of the full assessment centre with the combined stations. In this case, the corrected (for curtailment) test–retest correlation was 0.80, an estimate that may meet the standard for high-stakes decision making.[25–28] This correlation is still lower than the test–retest correlation found for cognitive ability tests like PET, which was about 0.90 (e.g. the correlation for 162 839 examinees repeating PET during 2000–2009 was 0.90 (A Allalouf, personal communication, 2010).

Limitations

The generalisability analysis presented in this study was based on a D study design. Unfortunately, as a result of logistical and cost considerations, no G study design that might have provided us with estimates of variance components (e.g. rater, SP, and various interactions of these variables with candidate) was feasible.

Another limitation of this study is that it did not take into account year-to-year changes that may have differential impact on candidate performance in each year, such as an increase in the number of candidates who participated in coaching programmes. An examination of data collected in the future will inform us of any systematic trend.

Furthermore, the impact of group composition on the results warrants investigation. The examinees represented several subgroups according to gender, native language (Hebrew or Arabic) and age (18–19 or 20 years old). It should be noted that the composition of the groups that participated in MOR and MIRKAM was fairly similar. As a result of practical constraints, analyses of the subgroups were beyond the scope of the present study. However, a detailed analysis of the results by gender, native language and age is currently in progress.

CONCLUSIONS

The only acceptable test–retest coefficient (0.80) was obtained for the full assessment centre, based on 14 stations as well as the two questionnaires. A test–retest correlation of 0.71 was obtained for the 14 stations. The closest approximation of the 'true' reliability of eight or nine MMI stations seemed to be 0.56. This estimate was obtained when the scores of the two MMI systems were correlated, and it was similar to the average of the test–retest correlations for the MOR and MIRKAM behavioural stations. More accurate measures of around 0.70 were obtained when the admissions system included additional measures such as the questionnaires.

It should be noted that although the above reliability estimates may appear somewhat lower than required for high-stakes selection decisions, they are superior to results obtained with a single pre-admission interview, even when the interview is well structured and involves more than one rater.[8,26] In light of these estimates, the justification for using these tools, which are logistically complex and demanding, should be discussed in terms of how much weight is assigned to them in the admissions process. In one of the schools, the new screening process resulted in changes of 18% and 20%, respectively, in the make-up of the cohorts admitted in 2004 and 2005. Each faculty will need to determine whether this is an acceptable result. The rejection of medical school candidate applications should be based on informed, consistent and reliable decisions. The MMI seems to be more reliable than a single interview. If resources were unlimited, it would be possible to double the number of stations and include questionnaires in order to achieve higher reliability. This, however, is not realistic. Nevertheless, the reliability of the system can be improved by thoroughly training the observers, interviewers and questionnaire raters. Moreover, it seems that the quality of evaluation is dependent upon rater attitudes and their recognition of the importance of what they do. The more involved the dean, members and chair of the admissions committee are and the stronger their trust in the process, the more likely they are to imbue evaluators with commitment and purpose.

Although investigating the reliability of the MMI admissions tool is important, gaining knowledge regarding its validity – the extent to which the candidate's performance, as evaluated by the MMI and the questionnaires, is related to his or her future professional behaviour – is even more crucial. Unfortunately, the data gathered thus far regarding the validity of the MMI are sparse and not yet satisfactory. Efforts are currently underway to collect information on predictive validity criteria from clinical evaluations of performance during clerkship in several hospital wards. In addition, a study that examines the relationship of assessment centre scores with burnout variables during and after medical studies is being conducted. A third study is examining the ability of MIRKAM and MOR (behavioural stations and questionnaires) scores to predict peer evaluations of various personal and behavioural dimensions during clerkship in Year 4 of medical studies.

The use of assessment centre behavioural stations for measuring non-cognitive attributes conveys an important message to the public, to the candidates, to participating raters and, by extension, to the health institutes they represent. It makes the point that in order to become a competent and ethically sensitive practitioner, it is necessary, but not sufficient, to be a good student. Attributes such as communication skills, motivation, integrity, maturity, self-confidence and social awareness are highly valued and essential.

Contributors: NG contributed to the study conception and design, the analysis and interpretation of data and the drafting and revision of the article. AM, OE, DZ and AZ contributed to the study conception and design and to the revision of the article. All authors approved the final manuscript for submission.
Acknowledgements: the authors thank Elliot Turvall, NITE, for prompt, efficient and professional data analysis, and Janine Woolfson, NITE, for copyediting and enlightening comments. The authors are also grateful to Ruth Beit Marom, (Open University of Israel, Raanana, Israel) Yoav Cohen, NITE, Avi Allalouf, NITE, Michal Beller, (RAMA, Tel Aviv, Israel), and the reviewers for their constructive remarks. The authors express their appreciation to contributing members of the National Institute for Testing and Evaluation and the Israel Centre for Medical Simulations, and to faculty members at Tel Aviv University and the Hebrew University for their assistance and input.
Funding: none.
Conflicts of interest: none.
Ethical approval: not required.

REFERENCES

1 Johnson ED, Edwards JC. Current practices in admission interviews at US medical schools. *Acad Med* 1991;**66**:408–12.
2 Edwards JC, Johnson ED, Molidor JB. The interview in the admission process. *Acad Med* 1990;**65**:167–75.

3 McGaghie WC. Student selection. In: Norman JR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht; Boston, MA; London: Kluwer Academic Publishers 2002; 303–35.
4 Ziv AM, Rubin OR, Moshinsky AV, Gafni NA, Kotler MO, Dagan YA, Lichtenberg DO, Mekori YO, Mittelman MO. MOR: a simulation-based assessment centre for evaluating the personal and interpersonal attributes of medical school candidates. *Med Educ* 2008;**42**:991–9.
5 Eva KW, Reiter IH, Rosenfeld J, Norman GR. An admissions OSCE: the multiple mini-interview. *Med Educ* 2004;**38**:314–26.
6 Lemay JF, Lockyer JM, Collin VT, Brownell AKW. Assessment of non-cognitive traits through the admissions multiple mini-interview. *Med Educ* 2007;**41**:572–9.
7 Harris S, Owen C. Discerning quality: using the multiple mini-interview in student selection for the Australian National University Medical School. *Med Educ* 2007;**41**:234–41.
8 Roberts C, Walton M, Rothnie I, Crossley J, Lyon P, Kumar K, Tiller D. Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Med Educ* 2008;**42**:396–404.
9 Reiter HI, Eva KW, Rosenfeld J, Norman GR. Multiple mini-interviews predict clerkship and licensing examination performance. *Med Educ* 2007;**41**:378–84.
10 Axelson RD, Kreiter CD. Rater and occasion impacts on the reliability of pre-admission assessments. *Med Educ* 2009;**43**:1198–202.
11 Eva KW, Reiter H, Rosenfeld J, Norman GR. The relationship between interviewers: characteristics and ratings assigned during a multiple mini-interview. *Acad Med* 2004;**79**:602–9.
12 Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions and tool development. *Adv Health Sci Educ Theory Pract* 2009;**14**:759–75.
13 Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. *Med Educ* 2009;**43**:767–75.
14 Eva KW, Reiter HI, Rosenfeld J, Norman GR. The ability of the multiple mini-interview to predict preclerkship performance in medical school. *Acad Med* 2004;**79** (Suppl):40–2.
15 Brennan RL. *Elements of Generalizability Theory*. Iowa City, IA: American College Testing Program 1986;55–70.
16 Brennan RL. *Generalizability Theory*. New York, NY: Springer 2001;95–136.
17 Beller M. Admission to higher education in Israel and the role of the psychometric entrance test: educational and political dilemmas. *Assess Educ Princ Pol Pract* 2001;**8** (3):315–37.
18 McLachlan JC, McHarg J. Ethical permission for the publication of routinely collected data. *Med Educ* 2005;**39**:944–8.
19 Wang M, Stanley J. Differential weighting: a review of methods and empirical studies. *Rev Educ Res* 1970;**40**:663–705.
20 Feldt L, Brennan R. Reliability. In: Linn R, ed. *Educational Measurement*, 3rd edn. New York, NY: Macmillan, American Council on Education 1989; 105–46.
21 Feldt L. Estimating the reliability of a test battery composite or a test score based on weighted item scoring. *Meas Eval Couns Dev* 2004;**37**:184–90.
22 Thorndike RL. *Applied Psychometrics*. Boston, MA: Houghton Mifflin 1976;208–15.
23 American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: AERA 1999.
24 Kreiter CD, Ping Y, Solow C, Brennan RL. Investigating the reliability of the medical school admissions interview. *Adv Health Sci Educ Theory Pract* 2004; **9**:147–59.
25 Buckendahl CW, Plake BS. Evaluating tests. In: Downing SM, ed. *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates 2006;725–38.
26 Schuwirth L, van der Vleuten C. The use of clinical simulations in assessment. *Med Educ* 2003;**37** (Suppl):65–71.
27 Nayer M. An overview of the objective structured clinical examination. *Physiother Can* 1993;**45** (3):171–8.
28 Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. *Fam Med* 2008;**40** (8):574–8.

Received 25 August 2010; editorial comments to authors 25 November 2010, 10 June 2011; accepted for publication 25 August 2011