ASSESSMENT FOR SELECTION FOR THE HEALTH CARE PROFESSIONS AND SPECIALTY TRAINING

14TH OTTAWA CONFERENCE ON THE ASSESSMENT OF COMPETENCE IN MEDICINE AND THE HEALTH CARE PROFESSIONS

15-20 MAY 2010

MIAMI

INDEX

ASSESSMENT FOR SELECTION FOR THE HEALTH CARE PROFESSIONS AND SPECIALTY TRAINING THEME GROUP

THEME LEADERS	Professor David Prideaux Flinders University, Adelaide, Australia Associate Professor Chris Roberts University of Sydney, Sydney, Australia
MEMBERS	Associate Professor Kevin Eva University of British Columbia, Vancouver, Canada Professor Angel Centeno University of Austral-Argentina, Buenos Aries, Argentina Professor Peter McCrorie St Georges, University of London, London, United Kingdom Professor Chris McManus University College London, London, United Kingdom Professor Fiona Patterson City University London, London, United Kingdom Professor David Powis University of Newcastle, Newcastle, Australia Associate Professor Ara Tekian University of Illinois, Chicago, USA Professor David Wilkinson University of Queensland, Brisbane, Australia

RECOMMENDATIONS

Recommendation 1. Admissions committees and all who have an interest in selection processes should adopt the principles of good assessment in defining the purpose of selection; blueprinting of assessable domains and attributes, selecting appropriate formats, employing transparent standard setting and decision-making, and including an evaluation cycle in a programmatic manner.

Recommendation 2. An integrative approach should apply the principles of good assessment and curriculum alignment along the education and training pathway including the progression hurdles between health professional degrees, prevocational practice and basic and advanced speciality training

Recommendation 3. There should be a focus on multi-method programmatic approaches in collecting, analysing, interpreting and reporting data from a range of selection instruments, which are fit for purpose.

Recommendation 4. There needs to be an emphasis on developing interdisciplinary theoretical frameworks that underpin development of both policy and the research agenda.

Recommendation 5. There is an urgent need for the development of sophisticated measurement models from the family of regression methods which will require application to multi-site high quality data sets, for increasing the sophistication of predictive validity studies using a range of attributes from selection blueprints, and for a focus on test-retest reliability.

Recommendation 6. The social accountability of universities demands that social inclusion, workforce issues, consumer choice and widening of access to students of promise are embedded in the principles of good assessment for selection with recognition that there are political (and non-universal) issues that need to be considered in the definition of optimal decisions.

Recommendation 7. Outreach, targeting strategies, preparation programs and conditional selection should be considered as core strategies for medical and health professional schools to achieve their widening access missions.

1. INTRODUCTION: ASSESSMENT FOR SELECTION

The term assessment is usually associated with the process of examining or testing candidates once admitted to a course or program. Yet the principles which underly assessment are every bit as important when applied up front to the very process of selecting those to be admitted to the course. Indeed, universally in the health professions, whether at initial or subsequent phases of training, there are many more applicants than places available. Careers in the health professions are satisfying and financially rewarding. Furthermore, attrition rates in health professional courses are low and once selected most entrants graduate. As a result, in common with much in-course assessment, the stakes can be high. Selection processes therefore need to be credible and fair, valid and reliable and, above all, publicly defensible. By conceptualising selection as 'assessment for selection' the well-developed quality assurance mechanisms associated with high stakes assessment can be applied to the selection process. These include:

- proceeding from a clear blueprint of the content for selection;
- using evidence from psychometric studies and a theory base to inform the selection process;
- developing congruity between selection, curriculum and assessment;
- using clear standard-setting and decision-making procedures;
- providing a focus on the impact of selection (a variant of the adage that assessment drives learning).

This paper summarises the current state of the findings on selection measures. Yet, despite the importance of selection, there are limitations in the literature. Like much of medical education research, few studies proceed from an explicit theoretical focus or even from a strong conceptual framework. Rather there is a concentration on the properties of individual selection measures. Most of the studies originate from North America where the graduate entry mode has resulted in greater emphasis on written tests and non-test selection measures. There is an emerging literature from the United Kingdom and Australia as new medical schools have opened in those countries, some of which have adopted graduate entry approaches. There is a small comparative literature from the Netherlands as alternatives are adopted to the national lottery system.

While the title of this paper includes reference to health care professionals and specialty training most of the literature concentrates on the selection of medical students for the initial phase of education. There is one study cited here on selection for Canadian Dental Schools[1] and, in Australia, the Graduate Australian Medical Schools Admission Test (GAMSAT) and Undergraduate Medical Admissions Test (UMAT) are used for selection to health professional programs such as dentistry, optometry and physiotherapy. Salvatori[2] has produced a review entitled *The reliability and validity of admissions tools used to select students for the health professions*. While the literature is drawn from across the health professions many of the findings of the review are drawn from medical education.

Additionally there are few studies of selection for postgraduate training and only three such studies are included in this review[3,4,5]. Selection at this level has not been subject to the same scrutiny as for initial medical education. The numbers of entrants are small and practices tend to be less openly discussed and investigated.

The literature, however, does fall into three categories of significance. The first two categories are commonly represented by a so-called cognitive, non-cognitive divide. The former consists of written tests and ratings of academic achievement such as grade point average (GPA) or equivalent. These are discussed in Section 2 and 3. The so-called non-cognitive measures include interviews and their OSCE-inspired variant, Multiple Mini-Interviews (MMI) and are discussed in Section 4-6. Nevertheless it is maintained that the cognitive/non-cognitive dichotomy is flawed and is not used further in this paper. While the constructs underlying OSCEs and MMIs are not always clear it is inconceivable that these forms of selection have no cognitive components.

The other area of considerable importance in selection is the question of widening access to medical and health professional courses to include greater representation of ethnic minorities, low socio-economic or disadvantaged groups or Indigenous peoples. This is a different dimension from that of the comparative merits of different types of

selection measures. It is a values not a technical question and has strong local and political dimensions. The term political validity is important here. Widening access is discussed in Section 7.

The final sections of the document provide a consensus statement and recommendations both for implementation and for further study to advance the understanding of this important endeavour in the medical and health professions.

2. THE CURRENT POSITION: WRITTEN TESTS

This section introduces another flawed dichotomy; that between aptitude and achievement tests. Aptitude tests purport to measure potential for achievement while achievement tests purport to measure actual achievement. The most well-known selection test is the North American Medical College Admission Test (MCAT). MCAT was developed out of the more general Scholastic Aptitude Test (SAT) in the post-Flexner reforms of North American medical education. Its use is now ubiquitous in that context. It contains four main sections; Physical Sciences, Verbal Reasoning, a Writing Sample and Biological Sciences. While the MCAT had its origins in an aptitude test at least the Physical and Biological Sciences sections purport to measure achievement. It is less clear for the remaining sections. The Graduate Australian Medical Schools Admission Test is now used in Australia and the United Kingdom. It is modelled on the MCAT and has three sections; Reasoning in the Humanities and Social Sciences, Written Communication and Reasoning in the Biological and Physical Sciences. It purports to be an achievement test at least for the first and third sections.

A second test is used in Australia for courses taking school leaver students; the Undergraduate Medical Schools Admission Test (UMAT). It has three sections; Logical Reasoning, Understanding People and Non-Verbal Reasoning. From 2006 some medical schools in the United Kingdom have used the United Kingdom Clinical Aptitude Test (UKCAT) which comprises measures of Verbal Reasoning, Quantitative Reasoning, Abstract Reasoning and Decision Analysis. These tests purport to measure potential or aptitude for medical study.

The predictive validity of MCAT is reasonably well established. Donnon et al[6] conducted a meta-analysis of 23 studies investigating the predictive validity of MCAT as it related to performance in medical schools and Step 1 of the United States Medical Licensing Examination (USMLE). They found a predictive validity coefficient of 0.39 for MCAT for performance in the preclinical years of the medical course and 0.6 for performance in Step 1 of the USMLE. The Biological Sciences section was the best predictor on both measures. Julian's[7] study involved two cohorts of entrants from 14 medical schools. Again the predictive validity of MCAT was measured against medical school and USMLE performance, the latter involving all three steps. Medical School performance was predicted best by combined MCAT and Grade Point Average (GPA) from prior degree studies. MCAT provided a substantial increment over prior GPA particularly in year 3 studies. MCAT scores provided superior prediction of scores in the steps of the USMLE. MCAT had the advantage of being a standard score while the determination of GPA scores varied by schools.

Studies by Koenig et al[8] and Tekian[9] report on the positive predictive validity of MCAT but with some caveats. Koenig et al found that MCAT scores were predictive of medical school and USMLE Step 1 performance and that there was no difference between prediction for men and women. The performance of three designated ethnic groups was 'overpredicted'. The study concluded that MCAT was not a perfect predictor and other variables such as 'diligence', 'motivation and 'communication skills' need further investigation. Tekian used a combination of medical school performance factors including withdrawal and graduation status, USMLE performance and students with 'significant events' to examine MCAT and prior GPA predictive validity with Under Represented Minority (URM) and non-URM groups. MCAT and prior GPA scores were correlated with success in medical schools but did not have sufficient ability to define or differentiate the success or failure of students considered 'at risk'.

The use of large scale testing outside of North America is a relatively recent phenomenon. The number of studies of the GAMSAT is small. As in the Northern American context Coates[11] found that GAMSAT and GPA scores were best predictors of medical school performance in year 1. He also demonstrated high levels of divergence between GAMSAT, initial GPA and interview scores in relation to year 1 performance. In a study of two Australian medical

schools Groves et al[12] found significant correlations with year 2 medical school assessments only for the Physical and Behavioural Sciences section of GAMSAT. There was a negative correlation with scores in clinical reasoning tests. More recently Wilkinson et al[13] found in a single medical school study that the School's selection criteria, prior GPA, interview and GAMSAT were only modestly predictive of performance in examinations in years 1 and 4 of a four year course .

No published studies of the UMAT were located for this paper. Lynch et al[10] in a study of two Scottish medical schools found that UKCAT did not predict performance in the first year of medical school. Further studies are required before definite statements can be made about the utility of this test.

While the properties of MCAT are well known there are simply not enough studies as yet to make conclusions about the other tests or the relative merits of so-called achievement or aptitude tests. Further not enough is known about the underlying constructs of the tests to confidently classify them as achievement or aptitude and whether it is even possible to do so. Much work remains to be done.

Much work also remains to be done on predictive validity. Most studies are focused on the relation between selection tests and in course assessment, 'tests predicting tests', and do not proceed from strong theoretical foundations. Not much is known about the relationships with other qualities on selection blueprints or indeed with eventual practice as a health professional. This requires more sophistication in measurement methods and in choice of outcome variables.

3. THE CURRENT POSITION: ACHIEVEMENT RATINGS

The academic achievement of potential candidates prior to selection is commonly incorporated into selection processes. In the North American graduate entry context this means the GPA achieved in the pre-selection degree. The predictive utility of GPA in combination with MCAT has already been presented as part of findings in the previous section. Kreiter and Kreiter's[14] meta-analysis has shown that there are positive correlations between initial GPA and subsequent performance. There is no clear evidence about the relative merits of GPA in science compared to non-science subjects. Didier et al[15] reported on a method to adjust GPA to equate for differences between institutions. The adjusted GPA showed improved relation to MCAT scores and better prediction of USMLE and medical school performance but only when large adjustments to institutional scores were necessary. Again the same limitations about predictive validity from a narrow range of potential selection blueprint attributes apply.

There are few studies of the utility of school leaving scores despite their widespread use internationally. McManus et al[16] found that A level grades for UK schools were predictive of medical career choice but the results of a general intelligence test were not. In a study from one school from the Netherlands the school leaving GPAs of students within the national lottery selection system were associated with shorter times for graduation, greater success in achieving preferred specialist training and greater scientific output[17]. In a study of Universities and Colleges Admission Scores (UCAS) in the United Kingdom Powis et al[18] found that higher scores were associated with being younger and male and related to ethnic origin and type of school. This has implications for discussions in the widening access section of this paper.

4. THE CURRENT POSITION: INTERVIEWS

The interview, face-to-face contact with single interviewer or a panel with varying degrees of structure, is a common part of selection processes. Despite its ubiquity there are very few studies defining its psychometric properties. Those that do exist do not indicate that the interview is a robust selection measure.

Kreiter et al[19] conducted a review of studies of interviewer reliability since 1990. Nine studies were reviewed in total. Reported reliability was varied widely which the authors attribute to the differing definitions of reliability. They concluded that there was not sufficient evidence to establish the reliability of interviews. Their own study investigated 92 applicants who were interviewed over two consecutive years having failed to gain a place in the medical school the first time around. Reliability was established using multivariant and univariant generalisability

theory. Their estimates of reliability ranged from 0.27 to 0.38 and they concluded that this range was not sufficient to establish the reliability of the interview in question.

In their review of the assessment of personal qualities for selection for medicine Albanese et al[20] reached a similar conclusion. They described the results of reliability and validity studies as 'equivocal'. Furthermore they indicated a high degree of variability amongst interview formats particularly the characteristics that they purport to measure. Stansfield and Kreiter[21] have indicated at least one way to improve reliability. In their study in one medical school they found higher reliability for ratings at the high or low ends of a rating scale rather than middle levels. As a result they argue that a three point ranking scale may be as useful as the commonly used five point scale.

A study for selection into orthopaedic residency[4] in one University indicates at least one of the potential problems of interview; interviewer bias. In a study of 135 single interviewers, it was found that clinician interviewers gave candidates more favourable rankings when personality preferences, as measured by the Myers-Briggs scale, matched, notably the dimensions for extrovert-sensing, sensing-thinking and sensing-judging. The matching did not occur for interviewers who were basic scientists or residents.

Meredith et al[22] found significant correlations of interview scores with clinical assessments but the best defence of the interview has come from a study of applicants to Canadian Dental Schools using the Canadian Dental Association structured interview[1]. The interview was the result of extensive work by the Association. The blueprint was based on a job analysis of dental work which defined eight essential competencies. Questions were based upon critical incidents which matched the competencies. All interviewers were trained and panels of two, faculty members and dental practitioners were used. An inter-rater reliability coefficient of 0.81 has been established over five cycles of admissions with 1467 applicants. In a study of 573 applicants to four schools in one year positive correlations were found between interview scores and years 3 and 4 clinical performance but academic performance was not predicted. The authors conclude that the interview has a place in the selection process along with the Dental Aptitude Test (DAT) and other measures. What this study perhaps indicates is that some psychometric properties of the interview can be improved provided sufficient attention is given to both its detailed construction and operation.

5. THE CURRENT POSITION: MULTIPLE MINI-INTERVIEWS (MMI)

The Multiple Mini-Interview (MMI) was first developed at the Michael G DeGroote School of Medicine at McMaster University in Canada. It applies the principles of the Objective Structured Clinical Examination (OSCE) to the interview context. The OSCE provides a series of short testing stations and has been shown to have superior reliability to the single long case. Similarly the MMI employs a series of short interview scenarios with a single rater in each station or scenario. Eva et al[23] indicate that, like the OSCE, the MMI overcomes the problem of poor test-retest reliability and context specificity where the measurement of an attribute in one context does not necessarily transfer to another. Test-retest reliability provides a better indication of the quality of a test than inter-rater reliability because it focuses on the overall test not just a component of its operation

Good predictive validity and reliability of the MMI have been established in studies by Eva et al [23,24,25], LeMay[26], Reiter et al[27] and Roberts et al[28]. Eva et al have found reliability coefficients of 0.78, 0.65 and 0.76 in three separate studies with a median reliability of 0.73 over eight administrations of a 12 station MMI at McMaster. They have also found significant correlations with the Canadian Qualifying Examination part II which employs an OSCE format for both a postgraduate and undergraduate sample[25]. The study by Reiter et al from the same institution involving 117 volunteers indicated that the MMI was a better predictor of success on clinical clerkship OSCE performance, clinical encounter cards and performance ratings than prior GPA and other measures of non-cognitive variables used in the admissions process. The MMI was also a better predictor of the ethical sections of Part I of the Canadian Qualifying Examination[27]. Importantly, these studies revealed a complementary relationship between MMI and GPA with GPA being more predictive of other academic outcomes such as progress test performance and the core rotations of the Qualifying Examination.

The study by Roberts et al[28] is drawn from the Australian context, the University of Sydney. The authors report a reliability coefficient of 0.7 on an eight station MMI for 485 candidates. There was a small but significant correlation with Section 1 of the GAMSAT, Reasoning in the Humanities and Social Sciences, but a small negative

correlation with Section 3 Reasoning in Biological and Physical Sciences. Interviewer subjectivity was responsible for majority of the measurement error and Roberts et al recommended increased rater training to address this. Roberts and his colleagues have also subjected 39 items in their bank to Item Response Theory (IRT) analysis[29] . No items were found to have differential item function (DIF) and the questions appeared to measure the unidimensional constant of 'entry-level reasoning skills in professionalism'.

Further studies have demonstrated other attributes of the MMI. The Eva et al study[23] demonstrated that increasing the number of stations had a greater impact on reliability than increasing interviewers. Dodson et al[30] demonstrated that reducing station length from eight to five minutes had little impact on reliability and it has also been shown the results of MMI appear not to be affected by security violations[31]. Kumar et al have provided some theoretical insights into how judges arrive at their decisions and the biases to which they are subject[32]. There is also evidence for both interviewer and candidate support of the process[32,33] and that, while the MMI may require more physical space, it requires fewer planning hours[34]. The MMI has been shown to have high reliability when based as part of selection for International Medical Graduates (IMG) into family medicine residencies[35] and gained support of candidates and interviewers when used as part of selection of Senior House Officers (SHO) for a regional paediatrics program[36].

6. THE CURRENT POSITION: OTHER MEASURES

Other measures used in the selection process include personal statements, autobiographical statements or letters of recommendation. However, there is no evidence that they are necessarily reliable or have predictive validity. In the Albanese review[20] of personal qualities in selection no research papers could be located on such measures nor could any evidence be found that they measured anything different from interviews.

Lievens et al[37,38] provide studies of a situational judgement test (SJT) based on written or video-based responses to hypothetical scenarios. Responses are selected from a list of alternatives. The method is used as part of a centralised selection system for medical and dental students in Belgium. The test showed greater predictive validity for GPA than science related and cognitive ability tests but only where curricula included specific interpersonal content.

There is growing interest in other measures. Carr[39] has investigated emotional intelligence (EI). However in the Australian context she found no correlation between measures of EI with UMAT or with Tertiary Entrance Rank (TER), the Australian equivalent of GPA for school leaving students. Based on negative correlations with a dysfunctional personality scale and final examinations Knights and Kennedy[40] argue for the use of such a scale in selection. Ziv et al[41] have produced evidence of reliability and content and face validity for a simulation-based selection test and Dore et al[42] have demonstrated correlations with MMI and clinical performance scores for computer-based responses to video-based scenarios.

There is also growing interest in the application of personality testing used in business or commerce careers for selection. Albanese et al's review[20] points to one of the difficulties with this approach. They point to Price et al's study[43] indicating 87 qualities of successful doctors. There is great variability in the qualities currently assessed through interviews, MMIs and other non-cognitive measures. The psychology literature has shown some acceptance of the 'big 5' personality characteristics Openness, Conscientiousness, Extrovertness, Agreeableness and Neuroticism but there have been few attempts to apply this to selection for the medical and health professions.

Powis et al[44,45] and Lumsden et al[46] have developed a Personal Qualities Assessment (PQA) tool comprising a mental agility test, a moral orientation scale and a NACE scale (Narcissim, Aloofness, Confidence and Empathy). It was administered to 507 volunteer applicants to Scottish medical schools in 2003 but played no part in the actual selection process. While good discrimination power was shown for the test it was not possible to correlate with other measures to determine predictive validity. It remains an interesting area for further development.

7. THE CURRENT SITUATION: WIDENING ACCESS

As indicated previously widening access is a values question not a technical question of choosing one selection method over another. Widening access is driven by socio-political concerns. These are real concerns. The competitive nature of entry to the medical and health professions has meant that certain groups within populations are not well represented in medical school cohorts. In the United States widening access means attracting more students from Under-Represented Minority groups (URM). In the United Kingdom there is a concern to attract socioeconomic diversity. In Australia, Canada, New Zealand and South Africa, amongst other countries, there is an interest in ensuring more Indigenous students and more rural students are represented in medical school populations. The latter is tied to concerns about workforce distribution as well as equity issues. It is acknowledged that rural students are more likely to practise in rural locations after graduation. A common approach has been to institute quotas for such groups.

There is increasing interest in the social accountability mandate of medical schools[47] including the formation of a Training for Health Equity network of medical schools (THEnet). Social accountability requires responsiveness to the communities the medical school serves and ensuring that the communities are represented in the student population. From this derives the concept of political validity. How does the selection process meet the requirements of the communities served by the schools and what impact does it have on the success of community members in gaining admission? The measurement of this form of validity will be informed not by a psychometric discourse but by one derived from the social and political sciences.

Certainly there is evidence from the United Kingdom to show that traditional methods of selection do discriminate against defined groups within the population. Powis et al[18] for example found that the UCAS tariff scores in the UK, the equivalent of a school leaving GPA, were associated with being young, male, less materially disadvantaged, white, Chinese or mixed ethnic origin and type of school attended. Similarly James et al[48] studied the new UKCAT for entrants to 23 UK medical and dental schools. They found that UKCAT scores could be used as a 'proxy' for the A levels school leaving examinations but there were biases towards male candidates and candidates from higher socioeconomic classes and independent schools. James et al[49] are advocates for graduate entry as a means to widen access. They compared school leaver and graduate applications to the Nottingham Medical School in 2002/3. The applicant pool for graduate entry contained more males, lower socioeconomic students and students with low UCAS scores than the school leaver programs. The differences were maintained in the actual entrants who gained places in the two programs.

Alternative selection processes aimed at widening access have not been extensively studied. Steinecke et al[50] discuss the evidence for four main types of approaches; traditional measures, socioeconomic measures based on personal or family demographic data, adversity indices based on assessment of current or prior disadvantage and community outreach strategies. Of the four approaches Steinecke et al conclude that socioeconomic measures hold 'promise' although requiring further studies. Traditional measures have not been successful in widening access and the outcomes of the use of adversity indices and outreach are as yet not well studied.

One of the problems faced in selecting for widening access is how to make decisions based on the data collected during selection and the socioeconomic data or measures of disadvantage. In 1997 Tekian[51] surveyed 15 medical schools who admitted URM students. He found that there were varied weightings for the various quantitative, qualitative and student disadvantage data used. He described the admission polices as 'secretive'. Technical solutions to this have been provided by Kreiter[52] (controlled optimisation), Bore et al[53] (multiple cut off scores in a regression model) and Reiter and Maccoon[54] (Hofstee standard setting). All claim that use of these methods can result in selection that increases student diversity.

One approach to widening access that is having some success is the development of special preparation programs through which members of under-represented groups can gain the skills to be competitive at entry. Reeves et al[55] have shown how a premedical certificate program has enabled previously rejected applicants to gain entry to the University of North Texas Health Sciences Center. Dalley et al[56] report on a statewide approach in Texas. The Joint Admissions Medical Program (JAMP) applies to all nine Texas medical schools. Applicants to the JAMP must produce evidence of economic disadvantage and have completed some University studies at first year level. They then follow an enrichment program throughout their studies including summer internships. The program has

proved popular with 1230 applicants in the first six years of operation and 164 of 288 participants gaining places in medical schools.

As indicated previously some schools have developed community outreach programs as part of their commitment to underserved populations in the communities they serve. These involve 'pipeline' and special preparation programs within comprehensive strategies of recruitment, retention and community service. Acosta and Olsen[57] report on the University of Washington's approach to attract and retain American Indian and Native Alaskan Physicians. They outline two recruitment programs involving attracting high school students into health courses and medical/dental summer schools. They combine this with retention, support and curriculum strategies once students are admitted. From 1989-2005 477 students have participated in the programs with 34 entering the University of Washington and 102 other medical schools.

Baylor College of Medicine has developed a partnership with the University of Texas-Pan American (UTPA), a largely Hispanic University[58,59]. The Preclinical Honors Program (PHC) involves an enriched program at the UTPA. Importantly conditional entry at Baylor is granted subject to GPA requirements being met. The program has resulted in an average of 12 students per year gaining entry since 1994 which represents a significant increase over previous figures. This community outreach and special program approach is being adopted elsewhere especially in Australia to recruit Indigenous Australians, many of whom reside in underserved communities and whose participation rate in health professional courses is low. These programs are not yet represented in the literature.

Outreach may well represent a promising development in addressing the question of widening access especially as more medical schools are embracing their social accountability mandates. Entry through a preparation program with conditional places overcomes the significant issue of combining different types of entry data. Combining selection with support and curriculum measures could well have the potential to add strength to this approach.

8. CONSENSUS AND CONCLUSION

8.1 Consensus

What the paper shows is that there is a consensus about assessment for selection for the health professions and specialty programs but the areas of consensus are small. There is evidence for the predictive validity of MCAT and GPA. There is not strong evidence as yet for the credibility of newer tests introduced in countries outside of North America such as GAMSAT, UMAT, or UKCAT. Nor is there much evidence outside North America about GPA of prior study whether it be in the form of high school leaving grades or prior university study. There is an obvious need for more studies in these areas.

For other measures there is evidence of the test-retest reliability and predictive validity of the MMI but not much else. Furthermore, there is evidence on this issue from outside North America. There is not much evidence of the credibility of interviews, personal statements and letters of reference.

8.2 Next steps

What was made clear from the outset in this paper that the study of selection for assessment is largely atheoretical. Only one article located in this review proceeded from a strong theoretical background. Sternberg[60] used the theory of intelligence to examine current approaches to selection, although the theory itself lacks current credibility. Interestingly, however, he draws many of his examples from outside of medical education. The discipline of medical education as a whole has been served well by looking outside the discipline's boundaries which are at best loose in any case. The adoption of outcomes-based curriculum from mainstream education is a good case in point.

Thus the study of selection measures could benefit from an examination from other literature, notably that in psychology and the social sciences. The study by Price et al[43] included 87 qualities that could be assessed for selection. Measurement of them all is not feasible. At the other end of the scale the 'big five' factors have been

identified for personal qualities in the psychology literature. David Powis and his colleagues[44,45,46] have developed Personal Quality Assessment drawing from the psychological and moral development literature. The area may prove to be of promise.

There are caveats. There may be dangers in moving to an overly psychometric view of the measurement of personal qualities in selection. Models of selection drawn from career selection and business and commerce may not transfer across into the professional world of health care providers. A view from across the social sciences not just psychology is necessary.

It may be useful to take a step backwards from the pursuit of unifying theory to consider Regehr's[61] concept of programmatic research where 'communities' of researchers work together towards an eventual goal of consensus. He claims research on the OSCE as one such example of programmatic research. This could also be applied to its selection-related variant, the MMI, where there is an emerging consensus about its credibility, feasibility and acceptability.

The development of the MMI has taken the key concept of context specificity and applied it to a recasting of the traditional interview into a new format just as the OSCE recast the traditional long case. In turn context specificity has a theoretical basis in social psychology and instance-based models of cognition. Are there other benefits to be gained from examining key concepts in other parts of our assessment processes and applying them to assessment for selection? van der Vleuten and Schuwirth[62] have argued that thinking about assessment should be moved from a consideration of methods to programs; another use of the term 'programmatic' this time in programmatic assessment. Programmatic assessment concentrates on the overall program of assessment with a combination of methods, each with their differing psychometric properties, to make decisions about student performance. This may prove profitable for medical schools in making decisions based on the variety of selection measures in use. There is perhaps one limitation in that the sheer numbers of applicants for places in medical and health professional programs would limit the number of measures that can be used. Nevertheless consideration of assessment for selection as a program rather than a collection of methods may well strengthen our understanding of this area.

Comparative studies of different programs of selection and their effects could provide important data for this endeavour. Given the high stakes nature of selection it is not practical nor defensible to divide applicants into different groups for selection for research purposes. However, there are opportunities for natural experimentation where alternative selection measures exist side by side such as in the Netherlands where individual schools are introducing alternative pathways for selection to the national lottery system[17,63]. There have been further studies of different tracks within individual medical schools[64,65].

Before concluding this section it is appropriate to turn to the title of this paper and again point out to the lack of evidence for selection outside of medical education and for specialty training. These too are areas for further activity and will bring new perspectives in the current debate. How does the selection of medical students differ from their health care colleagues particularly in environments where interprofessional team work is the norm? What is the effect of raising the stakes even higher as in specialty training?

8.3 Widening access

Adopting a programmatic approach to admission for selection may assist in one of the dilemmas in widening access; how to make decisions on the complex combination of merit, equity, fairness and social accountability issues represented on health professional selection blueprints. But it is also useful to remember that widening access is a values-based not a technically-based decision and hence it should be treated as such. Removing it from the technical domain should not be seen as downgrading its importance. Rather the reverse is true. It should be seen as a central point of a medical or health professional school's mission requiring different approaches to selection. This should be underpinned by the further development of the concept of political validity, a discourse derived from social sciences rather than psychometrics.

Widening access should not be seen as solely the domain of technical experts on the selection committee or medical education units. It requires the whole school to define its values position particularly with respect to the

communities it purports to serve. Where the communities are under-served or disadvantaged there are compelling reasons for schools to select and retain students from those communities.

The outreach approach combined with targeted enrichment or preparation programs and conditional selection has been outlined in the text of this paper. This approach is not new. It has some demonstrated success and deserves more attention as a fundamental part of assessment of selection.

8.4 Conclusion

A consensus statement on assessment for selection is important. Indeed conceptualising selection within an assessment framework gives it the possibility of additional intellectual scrutiny. The stakes for selection are every bit as high as for assessment for competence within a program, if not higher. Psychometric concepts like validity and reliability apply. Like in-course assessment there is a consensus on assessment for selection but only around a small number of methods. Again like in-course assessment, alternative methods are used, many without the evidence base that is desirable.

Keeping selection within the purview of assessment and applying concepts of programmatic selection will make it an effective area of study within medical and health professional education. Selection is, after all, the very first step in developing a competent and caring medical and health professional.

9. RECOMMENDATIONS

1. Admissions committees and all who have an interest in selection processes should adopt the principles of good assessment in defining the purpose of selection; blueprinting of assessable domains and attributes, selecting appropriate formats, employing transparent standard setting and decision-making, and including an evaluation cycle in a programmatic manner.

2. An integrative approach should apply the principles of good assessment and curriculum alignment along the education and training pathway including the progression hurdles between health professional degrees, prevocational practice and basic and advanced speciality training

3. There should be a focus on multi-method programmatic approaches in collecting, analysing, interpreting and reporting data from a range of selection instruments, which are fit for purpose.

4. There needs to be an emphasis on developing interdisciplinary theoretical frameworks that underpin development of both policy and the research agenda.

5. There is an urgent need for the development of sophisticated measurement models from the family of regression methods which will require application to multi-site high quality data sets, for increasing the sophistication of predictive validity studies using a range of attributes from selection blueprints, and for a focus on test-retest reliability.

6. The social accountability of universities demands that social inclusion, workforce issues, consumer choice and widening of access to students of promise are embedded in the principles of good assessment for selection with recognition that there are political (and non-universal) issues that need to be considered in the definition of optimal decisions.

7. Outreach, targeting strategies, preparation programs and conditional selection should be considered as core strategies for medical and health professional schools to achieve their widening access missions.

10. REFERENCES

1. Poole A, Catano V, Cunningham D. Predicting performance in Canadian dental schools: the new CDA structured interview, a new personality assessment and the DAT. *J Dent Ed* 2007; **71**, 5: 664-676.

2. Salvatori P. Reliability and validity of admission tools used to select students for the health professions. *Adv Health Sci Ed* 2001; **6**: 159-175.

3. Thordarson D, Ebramzadeh E, Sangiorgio S, Schall S, Patzakis M. Resident selection: how are we doing and why? *Clin Ortho and Rel Res* 2007; **459**: 255-259.

4. Quintero A, Segal S, King T, Black K. The personal interview assessing the potential for personality similarity to bias the selection of orthopaedic residents. *Acad Med* 2009; **84**: 1364-1372.

5. Hofmeister H, Lockyer J, Crutcher R. The multiple mini-interview for selection of international medical graduates into family medicine residency education. *Med Educ* 2009; **43**: 573-579.

6. Donnon T, Paolucci E, Violato C. The predictive validity of the MCAT for medical school performance and medical board licensing examinations: a meta-analysis of the published research. *Acad Med* 2007; **82**: 100-106.

7. Julian E. Validity of the Medical College Admission Test for predicting medical school performance. *Acad Med* 2005; **80**: 910-917.

8. Koenig J, Sireci S, Wiley A. Evaluating the predictive validity of MCAT scores across diverse applicant groups. *Acad Med* 1998; **73** 1095-1106.

9. Tekian A. Predicting performance and satisfaction: beyond the crystal ball. *Acad Med* 1998; **73**: 538-540.

10. Lynch B, Mackenzie R, Dowell J, Cleland J, Prescott G. Does the UKCAT predict year 1 performance in medical school? *Med Educ* 2009; **43**: 1203-1209.

11. Coates H. Establishing the criterion validity of the Graduate Medical Schools Admission Test (GAMSAT). *Med Educ*; **42** 999-1106.

12. Groves M, Gordon J, Ryan G. Entry tests for graduate medical programs: is it time to re-think? *Med J Aust* 2007; **186**: 120-123.

13. Wilkinson D, Zhang J, Byrne G, Luke H, Ozolins I, Parker M, Peterson R. Medical school selection criteria and the prediction of academic performance. *Med J Aust* 2000; **188**: 349-354.

14. Kreiter C, Kreiter Y. A validity generalization perspective on the ability of undergraduate GPA and the medical college admission test to predict important outcomes. *Teach and Learn in Med* 2007; **14**, 1: 95-108.

15. Didier T, Kreiter C, Buri R, Solow C. Investigating the utility of a GPA institutional adjustment index. *Adv Health Sci Ed* 2006; **11**: 145-153.

16. McManus I, Smithers E, Partridge P, Keeling A, Fleming P. A levels and intelligence as predictors of medical careers in UK doctors: 20 years prospective study. *BMJ* 2003; **327**: 139-142.

17. Cohen-Schotanus J, Muijtjens A, Reinders J, Agsteribbe J, van Rossum H, van der Vleuten C. The predictive validity of grade point average scores in a partial lottery medical school admission system. *Med Educ* 2006; **40**: 1012-1019.

18. Powis D, James D, Ferguson E. Demographic and socio-economic associations with academic attainment (UCAS tariff scores) in applicants to medical school. *Med Educ* 2007; **41**: 242-249.

19. Kreiter C, Yin P, Solow C, Brennan R. Investigating the reliability of the medical schools admissions interview. *Adv Health Sci Ed* 2004; **9**: 147-159.

20. Albanese M, Snow M, Skochelak S, Huggett K, Farrell P. Assessing personal qualities in medical school admissions. *Acad Med* 2003; **78**: 313-321.

21. Stansfield R, Kreiter C. Conditional reliability of ratings: extreme ratings are the most informative. *Med Educ* 2007; **41**: 32-38.

22. Meredith K, Dunlap M, Baker H. Subjective and objective admission factors as predictors of clinical clerkship performance. *J Med Educ* 1982; **57**: 743-751.

23. Eva K, Rosenfeld J, Reiter H, Norman G. An admissions OSCE: the multiple mini-interview. *Med Educ* 2004; **18**: 314-326.

24. Eva K, Reiter H, Rosenfeld J, Norman G. The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. *Acad Med* 2004; **79**: 602-609.

25. Eva K, Reiter H, Trinh K, Wasi P, Rosenfeld J, Norman G. Predictive validity of the multiple mini-interview for selecting medical trainees. *Med Educ* 2009; **43**: 767-775.

26. LeMay J-F, Lockyer J, Collin V, Brownwell K. Assessment of non-cognitive traits through the admissions multiple mini-interview. *Med Educ* 2007; **41**: 573-579.

27. Reiter H, Eva K, Rosenfeld J, Norman G. Multiple mini-interviews predict clerkship and licensing examination performance. *Med Educ* 2007; **41**: 378-384.

28. 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.

29. Roberts C, Zoanetti N, Rothnie I. Validating a multiple mini-interview question bank assessing entry-level reasoning skills in candidates for graduate-entry medicine and dentistry programmes. *Med Educ* 2009; **43**: 350-359.

30. Dodson M, Crotty B, Prideaux D, Carne R, Ward A, de Leeuw E. The multiple mini-interview: how long is enough? *Med Educ* 2009; **43**: 168-174.

31. Reiter H, Salvatori P, Rosenfeld J, Trinh K, Eva K. The effect of defined violations of test security on admissions outcomes using multiple mini-interviews. *Med Educ* 2006; **40**: 36-42.

32. Kumar K, Roberts C, Rothnie I, du Fresne C, Walton M. Experiences of the multiple mini-interview: a qualitative analysis. *Med Educ* 2009; **43**: 360-367.

33. Razack S, Fareno S, Drolet F, Snell L, Wiseman J, Pickering J. Multiple mini-interviews versus traditional interviews: stakeholder acceptability comparisons. *Med Educ* 2009; **43**: 993-1000.

34. Rosenfeld J, Reiter H, Trinh K, Eva K. A cost efficiency comparison between the multiple mini-interview and traditional admissions interviews. *Adv Health Sci Ed* 2008; **13**: 43-58.

35. Hofmeister M, Lockyer J, Crutcher R. The multiple mini-interview for selection of international medical graduates into family medicine residency education. *Med Educ* 2009; **43**: 573-579.

36. Humphrey S, Dowson S, Wall D, Diwakar V, Goodyear H. Multiple mini-interviews: opinions of candidates and interviewers. *Med Educ* 2008; **42**: 207-213.

37. Lievens F, Buyse T, Sackett P. The operational validity of a video-based situational judgement test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. *J Appl Psych* 2005; **90**, 3: 442-452.

38. Lievens F, Sackett P. Video-based versus written situational judgement tests: a comparison in terms of predictive validity. *J Appl Psych* 2006; **91**, 5: 1181-1188.

39. Carr S. Emotional intelligence in medical students: does it correlate with selection measures? *Med Educ* 2009; **43**: 1069-1077.

40. Knights J, Kennedy B. Medical school selection: impact of dysfunctional tendencies on academic performance. *Med Educ* 2007; **41**: 362-368.

41. Ziv A, Rubin O, Moshinsky A, Gafni N, Kotler M, Dagan Y, Lichtenberg D, Mekori Y, Mittleman M. MOR: a simulation-based assessment centre for evaluating the personal and interpersonal qualities of medical school candidates. *Med Educ* 2008; **42**: 991-998.

42. Dore K, Reiter H, Eva K, Krueger S, Scriven E, Siv E, Hilsden S, Thomas J, Norman G. Extending the interview to all medical school candidates – computer based-multiple sample evaluation of non-cognitive skills (CMSENS). *Acad Med* 2009; **84** (10 Suppl): S9-S12.

43. Price P, Lewis E, Loughmiller G, Nelson D, Murray S, Taylor C. Attributes of a good practising physician. *J Med Educ* 1971; **46**: 229-237.

44. Powis D. Personality testing in the context of selecting health professionals. *Med Teach* 2009; **31**: 1045-1046.

45. Powis D, Bore M, Munro D, Lumsden M. Development of the personal qualities assessment as a tool for selecting medical students. *J Ad and Cont Educ* 2005; **11**, 1: 3-14.

46. Lumsden M, Bore M, Millar K, Jack R, Powis D. Assessment of personal qualities in relation to admission to medical school. *Med Educ* 2005; **39**: 258-265.

47. Boelen C, Woollard B. Social accountability and accreditation: a new frontier for educational institutions. *Med Educ* 2009; **43**: 887-894.

48. James D, Yates J, Nicholson S. Comparison of A level and UKCAT performance on students applying to UK medical and dental schools in 2006: cohort study. *BMJ* 2010; **340**: 44.

49. James D, Ferguson E, Powis D, Symonds I, Yates J. Graduate entry to medicine: widening academic and socio-demographic access. *Med Educ* 2008; **42**: 294-300.

50. Steinecke A, Beaudreau J, Bletzinger R, Terrell C. Race-neutral admission approaches: challenges and opportunities for medical schools. *Acad Med* 2007; **82**: 117-126.

51. Tekian A. Minority students, affirmative action and the admission process: a survey of 15 medical schools. *Acad Med* 1998; **73**: 986-992.

52. Kreiter C. The use of constrained optimization to facilitate admission decisions. *Acad Med* 2002; **77**: 148-151.

53. Bore M, Munro D, Powis D. A comprehensive model for the selection of medical students. *Med Teach* 2009; **31**: 1066-1072.

54. Reiter H, Maccoon K. A compromise method to facilitate under-represented minority admissions to medical school. *Adv Health Sci Educ* 2007; **12**: 223-237.

55. Reeves R, Vishwanatha J, Yorlo T, Budd M, Sheedlo H. The post-baccalaureate premedical certification program at the University of North Texas Health Science Center strengthens admission qualifications for entrance into medical school. *Acad Med* 2008; **83**: 45-51.

56. Dalley B, Podawiltz A, Castro R, Fallon K, Kott M, Rabek J, Richardson J, Thomson W, Ferry P, Mabry B, Hermesmeyer P, Smith Q. The joint admission medical program: a statewide approach to expanding medical education and career opportunities for disadvantaged students. *Acad Med* 2009; **84**: 1373-1382.

57. Acosta D, Olsen P. Meeting the needs of regional minority groups: the University of Washington's programs to increase the American Indian and Alaskan Native physician workforce. *Acad Med* 2006; **81**: 863-870.

58. Thomson W, Ferry P, King J, Martinez-Wedig C, Michael L. Increasing access to medical education for students from medically underserved communities: one program's success. *Acad Med* 2003; **78** 454-459.

59. Thomson W, Ferry P, King J, Martinez-Wedig C, Villarreal G. A baccalaureate-MD program for students from medically underserved communities: 15-year outcomes. *Acad Med* 2010; **85**: 668-674.

60. Sternberg R. Assessing students for medical schools admissions: is it time for a new approach? *Acad Med* 2008; **83**: S105-110.

61. Regehr G. Trends in medical education research. *Acad Med* 2004; **79**: 939-947.

62. van der Vleuten C, Schuwirth L. Assessing professional competence: from methods to programmes. *Med Educ* 2005; **39**: 309-317.

63. Urlings-Strop L, Stijnen T, Themmen A, Splinter T. Selection of medical students: a controlled experiment. *Med Educ* 2009; **43**: 175-183.

64. Paolo A, Stites S, Bonaminio G, Cox G, Zeiger L, Meyer M, McCurdy S. A comparison of students from main and alternate admission lists at one school: the potential impact on study performance of increasing enrolment. *Acad Med* 2006; **81**: 837-841.

65. Hulsman R, van der Ende J, Oort F, Michels R, Casteelen G, Griffioen F. Effectiveness of selection in medical schools admissions; evaluation of outcomes among freshmen. *Med Educ* 2007; **41**: 369-377.