# Selection Methods for Foundation Programme: A Literature Review

**April 2009**

**Jan Illing**
**Madeline Campbell**
**Charlotte Kergon**
**Neill Thompson**
**Bryan Burford**
**Gill Morrow**
**Paul Crampton**
**Alex Haig**
**John Spencer**

NHS
North East Education
northerndeanery

## Acknowledgements

The team would like to offer thanks to our two research secretaries Tracy Straker and Elaine Knox who have worked tirelessly over the past seven weeks to support the team in putting this review together.

## Contents

## Glossary of Terms

| | |
|---|---|
| ABIM | American Board of Internal Medicine |
| ABOS | American Board of Orthopaedic Surgery |
| ABR | American Board of Radiology |
| ABS | Autobiographical Submission |
| ABSITE | American Board of Surgery In-Training Exam |
| ABSQE | American Board of Surgery Qualifying Exam |
| ACL | Adjective Check List |
| AEGD | Advanced Education in General Dentistry |
| AGCME | Accreditation Council on Graduate Medical Education |
| AMCAS | American Medical College Application Service |
| ANCOVA | Analysis of Co Variance |
| ANOVA | Analysis of Variance |
| AOA | Alpha Omega Alpha |
| ARS | Affect Reading Scale |
| ASF | Additional Selection Factor |
| BMAT | Biomedical Admissions Test |
| BOS | Behavioural Observation Scale |
| CARCS | Computer Assisted Resident Candidate Selection |
| CCEB | Canadian Chiropractic Examining Board |
| CDA | Canadian Dental Association |
| CDAT | Canadian Dental Aptitude Test |
| CPX | Clinical Practice Examination |
| CRA | Chief Resident Associate |
| CRP | Clinical Reasoning Problems |
| CSE | Clinical Skills Examination |
| DL | Dean’s letter |
| DTI | Diagnostic Thinking Inventory |
| DV | Dependant Variable |
| EEA | European Economic Area |
| EFL | English as a Foreign Language |
| EI tool | Emotional Intelligence tool |
| EM | Emergency Medicine |
| ERAS | Electronic Residency Application Service |
| ES | Educational Supervisors |
| FA | Factor Analysis |
| GAMSAT | Graduate Australian Medical School Admission Test |
| GPA | Grade Point Average |
| GPEP | Graduate and Professional Entry Program |
| GRE | Graduate Record Examination |
| HDS | Hogan Development Survey |
| IMG | International Medical Graduate |
| IQR | Interquartile Range |
| IRS | Interpersonal Reactivity Index |
| ISF | Interview Scoring Form |
| ISS | Interview Score Sheet |

| | |
|---|---|
| IV | Independent Variable |
| LMCC | Licensing Examination of the Medical Council of Canada |
| MCAT | Medical College Admission Test |
| MG | Medicine Clerkship Grade |
| MMI | Multiple Mini Interview |
| MOR | Selection for Medicine |
| MPL | Minimal Performance Levels |
| MRCGP | Member of Royal College of Practitioners |
| MSAT | Medical School Admissions Test |
| MSPE | Medical Student Performance Evaluation |
| NBME | National Board of Medical Examiners |
| NLSY | National Longitudinal Survey of Youth |
| NPTE | National Physical Therapy Examination |
| NRMP | National Resident Matching Programme |
| Obs & Gynae | Obstetrics and Gynaecology |
| OITE | The Orthopaedic In-training Examination scores |
| OMFS | Oral Maxillo-facial Surgery |
| OSCE | Objective Structured Clinical Examination |
| PASI | Portfolio Analysis Scoring Inventory |
| PBL | Problem Based Learning |
| PC | Personal Characteristics |
| PCA | Principal Components Analysis |
| PPQ | Patient Perception Questionnaire |
| PQA | Personal Qualities Assessment Test |
| PSQ | Patient Satisfaction Questionnaire |
| QST | Quantitative Composite Scoring Tool |
| REML | Restricted Maximum Likelihood |
| RL | Reference Letters |
| ROC | Received Operating Characteristic |
| RPQ | Reflective Portfolio Questionnaire |
| SC | Simulated Candidates |
| SD | Standard Deviation |
| SG | Surgery Clerkship Grade |
| SHO | Senior House Officer |
| SJT | Situational Judgement Tests |
| SLOR | Standard Letters of Recommendation |
| SP grades | Specialist Grades |
| SPR | Specialist Registrar |
| SPSQ | Standardized Patient Satisfaction Questionnaire |
| SRF | Standardised Recommendation Form |
| SSI | Social Skills Inventory |
| UGPA | Undergraduate Grade Point Average |
| UMAT | Undergraduate Medicine and Health Sciences Admission Test |
| USMLE | United States Medical Licensing Examination |
| VAS | Visual Analogue Scales |
| WPT | Wonderlic Personnel Test |
| WS | Weighted score |

# 1. Introduction

The NHS Next Stage Review states ‘Current arrangements for recruitment to Foundation Programme training are perceived by many in the [medical] profession as not robust enough. New work needs to be undertaken to develop more reliable and valid selection methods for recruitment to these programmes’ (Darzi, 2008, paragraph 32), and directs the newly established Medical Education England to consider recruitment processes. The Medical Schools Council, as a key stakeholder in the transition of medical students into practice, has commissioned this literature review to identify methods and processes used for selection which may inform future directions in recruitment to Foundation Programme.

## *1.1 Recruitment and selection to the UK Foundation Programme*

The Foundation Programme is young, having been introduced nationally only in 2005. One of its main initiatives was the introduction of a degree of centralisation in recruitment to postgraduate training posts which had previously been solely a matter for individual trainees and individual employers. With the introduction of the Foundation Programme, recruitment was through a national on-line application system, followed by local selection processes operated by Deanery-based Foundation Schools. The online Medical Training Application System (MTAS) required applicants to specify their preferred Foundation School, and give details of experience gained during their medical degree which were then rated by groups of assessors on a rating point scale, and an academic component indicated by their quartile ranking within their own medical school cohort. The Foundation School then may directly select high-scoring candidates, or have additional stages such as interviews or other assessments before selecting successful candidates.

The process is therefore a mixture of the national (with General Medical Council oversight of medical schools, and Department of Health operation of the MTAS system), and the local (Medical Schools having their own curricula and assessment methods, and Foundation Schools having discretion on the final stages of selection). This variation may provide strength in a heterogeneous population of doctors, and an ability to tailor recruitment to local workforce requirements, but has also led to the need for improvement identified by Lord Darzi. The current process contains risks that comparisons between different candidates may not be based on sufficiently standardised criteria, and that those criteria may not provide sufficient indications of the strengths and weaknesses of different doctors. While all new graduates who are eligible for provisional registration with the General Medical Council are de facto competent in the eyes of the law, they will have different levels of competence, and selection methods need to be sufficiently diagnostic to allow Foundation Schools to identify those who have a ‘best fit’ for their programme, and any areas of risk which may need to be addressed to get the best from those trainees, and *for* those trainees’ education.

## *1.2 Issues and methods in selection*

This review addresses a number of methods of selection. Many of these will be familiar, such as application forms, written examinations and tests, and interviews. However, these methods are not necessarily straightforward, and different types of interview for example will elicit different information, with different potential use. The choice of selection methods for Foundation Programme should be driven by a clear analysis of what is required for the job. In the same way as assessments should be ‘blueprinted’ against curriculum requirements (Crossley et al., 2002), effective selection requirements should be blueprinted, either by gaining expert views (e.g. Janis & Hatef, 2008) or conducting an appropriate job analysis (Patterson et al., 2000; Patterson et al., 2008).

Also, as with assessment, selection should be valid and reliable – that is, it should select on the basis of dimensions which are appropriate for the job to be done (blueprinting will help with this), and those dimensions should be consistent for all candidates. Marrin et al. (2004) found that those people involved in admissions (faculty, existing students and community representatives) valued validity and fairness as the most important elements of an admissions process. Reviews addressing issues of relevance and validity in a general, non-clinical, recruitment context are provided by Hough and Oswald (2000) and Sacket and Lievens (2008).

Many selection methods focus on the ‘non-cognitive’ elements of performance. This may include elements such as personality, communication and interpersonal skills, and other aptitudes which are not necessarily addressed by standard academic assessments. This is important for selection to Foundation Programme since, while it is likely that medical graduates fall into a relatively narrow range in terms of cognitive ability and knowledge base (a range restriction already present at entry to medical school; McManus, 2005), they are likely to have greater variance in terms of their wider, non-cognitive skills. As well as aptitude to do a job, selection may also have a role in identifying those likely to *stay* in a job (Barrick & Zimmerman, 2005).

Methods are rarely used in isolation. Application forms, interviews, and other methods are used together to form a selection profile. In some cases this involves the review and integration of disparate evidence by a recruitment panel, while in others quantitative data may be integrated by a set formula leading to explicit ranking. An approach increasing in popularity is the use of ‘assessment centres’, whereby a battery of assessments are completed in one place, at one time. Some approaches to integration of disparate evidence use complex weighting algorithms to relate retrospective data to prospective performance (e.g. Kretier & Solow 2002, Suhayda et al., 2008, Hemaida & Kalb 2001, Pilon & Tandberg 1997).

Selection should also be sensitive to issues of diversity and equal opportunity. While for UK graduates many issues are better addressed at the point of selection to medical school (e.g. McManus, 1998), the Foundation Programme is also open to applicants from overseas (Trewby, 2005 [although pre-FP]), and selection methods must be fair and unbiased, in terms of gender, ethnicity or socioeconomic background. Dewhurst et al. (2007) found that ethnicity was a factor in outcomes on the Membership exam for the Royal College of General Practitioners. Statistical treatments of retrospective application data may allow diversity to be addressed without having to resort to explicit ‘positive discrimination’ (Kreiter et al., 2003). Other strategies may involve training recruitment or admissions staff on non-cognitive selection methods (Price et al., 2008).

### *1.3 Aims of the review*

The review aimed to look at the published literature with the intention of identifying a) what methods of selection have been used in clinical contexts, b) what relevance they may have for selection to the UK Foundation Programme.

## 2. Method

The review involved three main stages: database search, filtering by abstract, and detailed review.

1. A comprehensive literature search was conducted across the most relevant databases (Medline, Embase, Psychinfo and ERIC) in accordance with the guidance developed for the Best Evidence Medical Education systematic reviews. The search strategies created were designed for maximum sensitivity (recall) to ensure that all efforts were taken not to overlook any papers of significance. The searches covered the last twenty years of research and did not limit results by geography, language or study methodology.

   Scoping strategies for each of the four databases were drawn up using the specific controlled vocabularies. Key papers of core relevance to the topic were then identified and the strategies were refined using combinations of controlled vocabularies, free text and search syntax. Strategies were finalised when they retrieved all the key papers known to be in each database. Initial full abstract lists were then visually scanned by the Information Scientist and clearly false hits were eliminated.

   Medline produced the largest number of results (8468) of which 762 were judged to have abstracts of potential relevance to the topic. The full Medline strategy is included as Appendix A. Embase returned 4697 hits of which 134 remained after duplicates with Medline were eliminated. Psychinfo produced 477, of which only 14 were close matches and not duplicates. ERIC returned no further relevant results after duplicates were removed. The total number of citations across the databases was therefore 910.

   Searching for evidence on medical education topics is well known to be more difficult than searching on entirely medical or entirely educational topics. A comprehensive /sensitive strategy is nearly impossible for medical education topics without compromising specificity (precision); this group, compared to BEME systematic reviews, was faced with a typically large number of false hits in order to avoid missing key evidence.

2. Filtering by abstract. Abstracts for all 910 papers (where the abstract was included in the database results) were read by a member of the research team, and considered against the inclusion criteria:

   - Clinical domain
   - Selection
   - Primary data
   - English language

- Peer reviewed article

If a paper was felt to satisfy the criteria, or in cases where an abstract was not available but the title suggested it might do, the full paper was obtained from electronic journals, library hard copies, or inter-library loan. This reduced the total number of papers to 359. Papers which did not meet all the criteria, but were nonetheless of interest – for example review articles, or articles from non-clinical domains – were also obtained.

3. Detailed review. Each of the obtained papers was read by a member of the research team, and content recorded on a pro-forma summarising the key points: aim, participants, design, results and conclusion, as well as any other notes. If a paper was not felt to be relevant following this review, a note was also made to this effect. Review papers, comments and editorials were not reviewed in detail, but were read to identify any salient points. The final number of papers reviewed in detail was 190.

## 3. Results: Overview

The review identified a range of studies looking at different elements of selection. Some areas found more literature, for example there were considerably more studies on interviews compared to on ‘white space’ questions.

Some selection methods, of which interviews are the most obvious, are prospective. That is, they involve an active intervention on the part of the recruiting body to measure or assess dimensions. Others are retrospective, looking at prior evidence of skills or qualities in the form of academic performance or extra-mural activity. These retrospective methods may include a candidate’s ‘grade point average’ indicating their overall undergraduate academic performance, or letters of recommendation/references from supervisors.

Many of the studies identified consider the effectiveness of different methods in terms of their relationship with later clinical or professional performance (a form of ‘predictive validity’ or ‘criterion validity’). In these cases attention should be paid to how specific findings may be. Caution should also be exercised where studies may have described a ‘predictive’ relationship, when simple correlations may have been described. These studies demonstrate a relationship between a measure and an outcome (e.g. interview scores and clinical performance) but they do not control for the effect of other measures (e.g. exam scores). Studies reporting regression results (rather than just correlation), typically assess the value of one measure in predicting performance, over and above what other measures predict (incremental predictive validity), e.g. an interview may be related to clinical performance, but may not predict clinical performance if exam scores are also used. Jargon has been avoided where possible.

The review is structured as follows: each broad class of method is presented as a self-contained section, with any distinct types indicated by sub-heading, and a concluding summary of the issues raised and any common trend in findings. Within each class, there is much variation in method, so disparity in findings of different studies should not necessarily be taken to indicate inconsistency in that method per se. Many studies describe multiple methods, used in aggregate or sequentially. These are included only once, so although each section is free-standing, there are cross-overs between the methods.

## 4. Interviews

The employment interview has been and remains the most widely used method of personnel selection (Barclay, 1999; Garman & Lesowitz, 2005). In medicine, Nayer (1992) reported that interviews are part of the admissions process in 99% of US medical schools and surveys of programme directors indicate that they place considerable weight on interviews when making admission decisions (e.g. Johnson & Edwards, 1991). Interviews are often included in the selection process as an opportunity to assess the non-cognitive attributes of applicants (Otero et al., 2006), which are regarded as critical to success as a medic, as reflected by their inclusion in the accreditation guidelines of the General Medical Council in the UK, the American Board of Medical Specialties, the Accreditation Council for Graduate Medical Education in the US, and CanMEDS 2000 in Canada (Harris & Owen, 2007). Historically, interviews have been regarded as invalid, unreliable predictors of job performance prone to several biases (Guion, 1998), but more recent developments have led to considerable improvements in the validity of this enduring selection procedure.

Interviews broadly include any recruitment measure which involves the oral questioning of a candidate. There is considerable variation in how interviews are conducted: one-to-one or with a panel, tightly structured and standardised or entirely at the discretion of the interviewer, brief interactions or long and detailed question and answer sessions, with highly trained and skilful interviewers or untrained novices, and face-to-face or over the telephone or via videoconferencing. In addition, the purpose of the interview may vary, with some organisations viewing them as a recruitment exercise and others focusing on assessment of candidates against pre-determined competencies.

Structure in interviews is one of the most critical moderators of its success in the prediction of performance (e.g. Huffcutt & Arthur, 1994). Research has consistently demonstrated that structured interviews typically have higher predictive and concurrent validity than unstructured interviews (e.g. McDaniel, Whetzel, Schmidt, & Maurer, 1994; Wiesner & Cronshaw, 1988). However, there is some ambiguity over how structure is conceptualised and operationalised. Campion, Palmer, & Campion (1997) defined interview structure as "any enhancement of the interview that is intended to increase standardization or otherwise [assist] the interviewer in determining what questions to ask or how to evaluate responses" (p.656); and Huffcutt and Arthur (1994) described it as "the reduction in procedural variance across applicants, which can translate into the degree of discretion that an interviewer is allowed in conducting the interview" (p.186).

Despite evidence of their low reliability and validity, unstructured interviews are still used for selection purposes. In a recent survey of pharmacy colleges in the US, only 13% of those that interviewed applicants used a pre-defined list of questions (Joyner et al., 2007). Unstructured interviews do not use a standardised set of questions, do not have objective scoring protocols (Latif et al., 2004), and are often prone to biases that affect ratings (Patterson & Ferguson, 2007). For example, halo effects (where one positive characteristic of an applicant influences ratings of their other attributes), horn effects (where one negative characteristic of an applicant influences ratings of their other attributes), similar to me effects (applicants are rated more favourably if they are similar to the interviewers with respect to demographics/professional interests), leniency (interviewers tend to use the upper end of the rating scale), stringency (interviewers tend to use the lower end of the rating scale), first impressions, and stereotyping (Joyner et al., 2007; Patterson & Ferguson, 2007). Another important disadvantage of unstructured interviews is the lack of legal defensibility (Posthuma et al., 2002).

Patterson and Ferguson (2007) described key features of a structured interview: 1) questions are directly related to the person specification (which is based on a thorough job analysis); 2) the same questions are asked of each candidate and follow-up questions may be used to elicit evidence but prompting is limited; 3) questions are designed to be situational, competency-based, biographical, or knowledge-based, and should be relevant; 4) interviews are longer or include a larger number of questions; 5) minimise the input of ancillary information (e.g. CVs, application forms, test scores); 6) only allow questions from the candidates after the interview, once ratings have been made; 7) score each response individually using standardised rating scales; 8) use detailed anchored rating scales and take notes to justify rating (based on evidence, not inference; 9) use multiple interviewers where possible (whilst still ensuring efficiency); and 10) train interviewers extensively to increase reliability. Empirical and theoretical support for these elements is described in Campion et al. (1997) and structured interviews have a greater degree of legal defensibility if a selection system is challenged (Posthuma et al., 2002).

Two types of questions are frequently used in structured interviews, in addition to questions asking about specific knowledge and biographical experience: competency-based and situational. Competency-based interview (CBI) questions ask candidates to give examples of their behaviours and actions in a specific situation (i.e. tell me about a time when…) and are based on the notion that the best predictor of future performance is past behaviour. Situational interviews (SI) present hypothetical situations and ask candidates to describe what actions they would take (i.e. what would you do if…?). Outside of medicine, meta-analytic evidence suggests that both CBI and SI are able to predict job performance (Huffcutt et al., 2004; McDaniel et al., 1994), although Huffcutt et al. (2004) found that SI validity was moderated by job complexity (such that validity decreased as job complexity increased), but CBI validity was not affected by job complexity.

This chapter will review studies on interviews that investigate issues relating to the reliability, validity, user reactions, and implementation of interviews. They will be summarised in discussion sections, alongside some additional research from non-medical fields. A final conclusion offers an overview of the findings.

### *4.1 Studies investigating the reliability of interviews*

O'Neill et al. (2009) examined the reliability of components of the selection process at a medical school in Denmark. Applicants (n = 307) took part in a 25 minute semi-structured interview to assess a range of attributes (e.g. subject interest, social skills, stress tolerance), scored by two independent raters. They reported good reliability for the interview (generalisability coefficient = 0.86), and the effect of raters was found to be small.

Oosterveld and Cate (2004) investigated the reliability of an interview for selection into medical school in The Netherlands. A 20-minute interview, described as structured (but little information was given) was conducted by a panel of three trained interviewers. Generalisability coefficients showed good reliability (ranging from 0.74 to 0.83). Using generalisability theory, the authors predicted that reliability could drop to 0.60 with one rater (instead of a panel of three).

Courneya et al. (2005) compared the reliability of a structured panel interview with an unstructured individual interview for selection into a Canadian medical school. Candidates (n = 33) were interviewed four times: 2 x unstructured individual interviews and 2 x 3-member structured panel interviews. In the unstructured interviews, there were no standard questions and candidates were scored on eight standard criteria and one global ranking. In the structured interviews, standardised questions assessed key attributes (e.g. integrity), which were rated using a behaviourally-anchored rating scale. Inter-interviewer reliability for the unstructured interview was very low at 0.12, and inter-panel reliability for the structured interview was 0.52, although the panel interviewers had received more training. Interview scores assigned by each panel were significantly correlated (ranging from 0.42 for interpersonal skills to 0.64 for motivation). There was consistency within panels, with scores from all three members (representing clinicians, academics, and individuals from the community) predicting the final consensus score. There were no significant correlations between panel interview score and academic indices (e.g. GPA, MCAT), or between scores from the unstructured and structured interviews.

Collins et al. (1995) studied the selection process in a medical school in New Zealand (see ‘studies investigating the validity of interviews’ section for details). Candidates (n = 79) were interviewed twice – each time by a panel of two trained interviewers using a structured protocol. Each panel member independently rated the candidate then developed a consensus score. Scores in the two panel interviews were correlated (r = 0.67), although reliability indices were not calculated.

Poole et al. (2007) examined relationships between predictors, including a structured interview, and the clinical and academic performance of dental students in Canada (see ‘studies investigating the validity of interviews’ section for details). Two trained interviewers rated candidates on seven competencies, with inter-rater reliability of 0.67, although the same interview protocol has previously demonstrated higher inter-rater reliability (0.81).

Gilbart et al. (2001) assessed reliability of an interview rating form based on 9 dimensions essential for successful performance in an orthopaedic surgery residency, and used in the selection of 66 candidates across 12 different orthopaedic surgical residency programmes in Canada. Candidates were scored on each dimension and given a score for ‘overall impression’. Interviewers independently rated each candidate, ranked all candidates, then final rank order was established by consensus. Interview format varied across institutions (structured/unstructured, no. of interviewers, duration). The internal consistency of the 9 dimensions was moderately high at Cronbach’s alpha = 0.71. Inter-rater reliabilities (within panels, across raters) were low to moderate, but increased (from ICC = 0.45 to 0.67) as assessment became more global, with greater differentiation of candidates on global measures. Mean ICCs within programmes/across panels were 0.58 for the sum of 9 items, 0.63 for overall impression score, and 0.63 for individual interviewer rank order; indicating moderate reliability in the ratings across interview panels in the same programme. Interestingly, no significant differences were found between reliabilities for programmes using structured versus unstructured interviews, although this finding is inconsistent with many other studies. Across different programmes, there were low correlations between scores given to the same candidate (ranging from 0.14 for the consensus final rank order to 0.18 for the sum of 9 items and overall impression score). Results demonstrate low to moderate reliability between raters in the same panel, and moderate reliability across panels in the same programme. Ratings of the same candidate across different programmes were inconsistent, suggesting there is considerable variation between programmes in how they score candidates.

Patrick et al. (2001) assessed the reliability and validity of a new structured interview for admission to a medical school residency programme (see ‘studies investigating the validity of interviews’ section for details). They reported good inter-rater agreement for the interview (2-person, 15-minute panel interview + 10-minutes to score using a behaviourally-anchored rating form): the percentage of rater pairs whose scores differed by one point or less ranged from 82% to 98%.

Harasym et al. (1996) assessed the inter-interviewer reliability and validity of interviewers' judgements of non-cognitive traits in medical school applicants (using simulated actor-candidates; see 'studies investigating the validity of interviews' section for details). Findings showed that interviewers were accurate in 56% of interviews. Good candidates were identified 67% of the time, average candidates 41%, and poor candidates 58%. Generalisability analysis found a large portion of variance (45%) in rating from one interviewer to the next. This variability is reflected in the generalisability coefficient of 0.51, which indicated only moderate inter-interviewer consistency in rating. Experienced interviewers were significantly more accurate (correctly rated more simulated candidates) than novice interviewers. As an aggregate, experienced interviewers correctly rated 70% of the interviews, whereas novice interviewers correctly rated 31% of the interviews.

### *4.2 Discussion: Interview Reliability*

Research in the medical literature suggests that interviews can be moderately reliable, but there is considerable variability in reported reliabilities, and several examples of low reliability exist. Oosterveld and Ten Cate (2004) reported generalisability coefficients of between 0.74 and 0.83 for a 'structured' interview (although few details were given on the interview format), indicating fairly good reliability, but they also calculated that reliability could drop to 0.60 if they had used only one rater instead of three. O'Neill et al. (2009) found good reliability for a semi-structured admissions interview (G = 0.86). Poole et al. (2007) reported that a structured interview conducted by two trained interviewers had demonstrated a good inter-rater reliability of 0.81, but found that reliability dropped to 0.67 in their study, despite using the same interview protocol. Patrick et al. (2001) found good inter-rater agreement with a panel of two interviewers using a structured protocol, with between 82% and 98% of rater pairs assigning points that differed by one point or less (out of 5). However, Courneya et al. (2005) reported a very low inter-interviewer reliability of 0.12 for an unstructured interview and a higher inter-panel reliability of 0.52 for a structured interview. Although this demonstrated that adding structure to the interview improves its reliability, it also indicated that high reliability is not necessarily guaranteed by using a structured interview format. Gilbart et al. (2001) found low to moderate inter-rater reliability using a standardised rating form (ICC = 0.45-0.67) and moderate reliability across panels in the same programme (ICC = 0.58-0.63), but reported low correlations between interview scores and ranks assigned to the same candidates by different programmes (r = 0.14-0.18). Harasym et al. (1996) reported moderate inter-interviewer reliability overall (G = 0.51) but found that experienced interviewers could score candidates more accurately than novice interviewers.

Studies in non-medical fields, such as occupational psychology, have also investigated interview reliability. In general, research also indicates that structuring interviews can improve reliability and that incorporating key elements of structure can improve reliability, as well as validity. In a widely cited review, Campion et al. (1997) proposed that increasing the standardisation of an interview with respect to questions, probes, rating and scoring protocols, ancillary information, interviewer training, and note-taking to support ratings, as well as using multiple interviewers, should improve reliability, as well as validity.

### *4.3 Studies investigating the validity of interviews*

Goho & Blackman (2006) examined the predictive validity of interviews for academic and clinical performance in healthcare disciplines using a meta-analytic approach. Twenty studies were included in the analysis: 19 of them reported on academic performance (total n = 4629) and 10 reported on clinical performance (total n = 1283). They found a mean weighted r=0.06 (95% confidence intervals: 0.03-0.08) for the predictive validity of interview performance in predicting academic success, indicating a very small effect of little practical value. They also reported a mean weighted r=0.17 for the validity of interview performance in predicting clinical performance, indicating a modest positive effect. The authors concluded that selection interviews do not effectively predict academic performance in healthcare disciplines, but have a modest capacity to predict clinical performance, although the effect size suggests that they may be of limited practical value.

Altmaier et al. (1992) compared the predictive validity of a competency-based interview (CBI) and traditional selection interviews on different aspects of performance in US radiology residents four years later (n = 72, 48% participation rate). Regression analyses indicated that scores on the CBI added incremental predictive validity over and above unstructured traditional interviews. Traditional interviews together with research and publications predicted only 2% of the variance in one performance measure (a behavioural observation scale) but 26% of the variance was explained with the addition of the CBI. Faculty ratings from traditional interviews negatively predicted performance, whereas a component of the CBI positively predicted performance. Traditional interviews together with research and publications were also found to predict 4% of the variance in another performance measure (director's evaluation form), which increased to 26% with the addition of the CBI. These results indicate that the CBI significantly added predictive validity to the selection process, whereas the unstructured traditional interview did not positively predict performance. Rather, higher scores on the faculty unstructured interview predicted poorer performance on the behavioural observation scale.

Olawaiye et al. (2006) carried out a study to assess the predictive validity of a residency selection process for clinical performance at the end of the 1st postgraduate year. Candidates (n = 107) were interviewed, scored independently, and divided into 2 groups based on rank percentile: the top half (50-99%) and the lower half (0-49%). There was a significant correlation between rank percentile and total performance score (r = 0.60, p, 0.001). Of those residents ranked in the top half group at selection, 12 out of 12 had a total performance score above 6 (out of 9). Only one out of 14 residents in the lower half had a total performance score above 6. The mean total performance score was higher and had significantly less variance in the top half group (mean=6.4, SD = 0.21) than the lower half group (mean= 5.3, SD =0.98). However, some of the attending physicians that evaluated residency performance may have also interviewed candidates (although these ratings were 18 months apart) and interviewers had access to application information on the candidates' cognitive attributes, which may have influenced their interview ranking.

Hoad-Reddick et al. (1999) assessed the predictive validity of admissions interviews at a UK dental school. Candidates (n = 102) were rated by a panel of 3 trained interviewers on 6 criteria (e.g. professionalism, leadership communication skills). Results showed that, during year 1, students with higher interview scores for leadership were less likely to fail (odds ratio: 0.46, 95% CI: 0.22-0.98). None of the interview scores predicted failure during year 2 of dental school.

Kinicki et al. (1990) assessed the relative effects of resume (CV) facts and interview observations on the selection of nurses (n = 312) at a US hospital, and their predictive validity for performance, organisational commitment and job satisfaction. Cues from the interview tended to dominate hiring decisions over resume credentials, but individual interviewers varied in how they used interview cues to make hiring decisions and no single interviewer predicted job performance. Interviews significantly explained 11% of the variance in organisational commitment, and 13% of the variance in job satisfaction, but no single interviewer positively predicted these job attitudes, and one interviewer negatively predicted attitudes. This study provides some evidence for the predictive validity of aggregate-level interview ratings (across 4 interviewers) in predicting job performance, organisational commitment and job satisfaction.

Basco et al. (2000) investigated the predictive validity of selection profiles from interview scores, as well as academic profiles (the grade point ratio and Medical College Admission Test scores), for third year OSCE performance of 222 medical students entering a US medical school. No significant correlations were found between rankings based on interview scores or academic profiles and OSCE performance. The study concluded that neither interviewer nor academic profiles predict performance on OSCEs.

Basco et al. (2008) carried out a study to test whether admissions interviews predict interpersonal interactions between final year medical students (n = 221) and standardised patients at a US medical school. Interview scores were based on three one-on-one interviews to assess how well the applicant fits an ‘ideal physician’ profile (e.g. interpersonal skills, judgement, empathy). Results indicated that interviews had a weak but significant correlation with OSCE performance (r = 0.15, p<0.05 for ratings of interpersonal skills; r = 0.13, p = 0.56 for overall performance). This suggests that admission interviews capture elements that have effects in clinical practice, but this small effect has limited predictive validity.

Heintze et al. (2004) examined the relationships between a range of selection methods, including interviews, and performance at a dental school in Sweden. Dental students (depending on method, n = 74 to 191) were rated by trained interviewers who did not have access to any additional candidate information. Interview scores were negatively correlated with the proportion of courses failed (r = -0.25, p<0.05), but did not predict the proportion of courses failed over and above spatial ability tests. Interview scores were not related to results on a pre-clinical course in operative dentistry or to course drop-out/breaks in study/repetition of course work. However, interview scores were positively related to interpersonal competence in meetings with their supervisor (r = 0.31, p<0.05) and professionalism in patient interaction (r = 0.30, p<0.05), as rated by clinical supervisors and tutors during the course. Regression analyses found that all possible combinations of predictor variables failed to account for significant levels of incremental variance in social competence and empathy over and above a single predictor. These results suggest that interview scores are related to course failure and ratings of social competence and empathy in dental school, but that they may not predict incremental variance, over and above other admissions tools (ability tests, etc). However, these predictors were not actually used for selection, and students completing tests and interviews may not have taken them as seriously as if they had real consequences.

Brothers and Wetherholt (2007) examined the relationships between selection criteria for 26 surgical residents and subsequent performance at a US medical school. Personal characteristics scored at interview (e.g. motivation, professional integrity) were more strongly related to clinical performance (with correlations ranging from r = 0.21 with professionalism, to 0.46 with system-based learning) than were exam scores and GPA, which were negatively related to clinical performance. Frequency of residents being identified as a ‘cause for concern’ was also negatively related to personal characteristics (r = -0.55). However, personal characteristics demonstrated no relationship with ABSITE scores and a weak negative relationship with ABS QE scores. This study indicated that non-cognitive personal characteristics assessed at interview correlated with clinical performance, although relationships may be inflated as the surgical faculty provided both interview scores and clinical performance scores.

Poole et al. (2007) examined the relationships between predictors, including scores on a carefully constructed competency-based structured interview, and the clinical and academic performance of dental students in Canada (n = 373). Two trained interviewers rated candidates on seven competencies, with inter-rater reliability of 0.67, although the same interview protocol has previously demonstrated higher inter-rater reliability (0.81). Correlational analyses indicated that the structured interview was significantly positively related to 3rd and 4th year clinical performance ($r = 0.18$, $p<0.05$; and 0.25, $p<0.001$, respectively), but not to other performance measures. When the correlations are corrected for range restriction and unreliability they rise to $\rho = 0.31$ for 3rd year and $\rho = 0.44$ for 4th year clinical performance. Interview scores also correlated with openness to experience ($r = 0.19$, $p<0.001$) and extraversion ($r = 0.26$, $p<0.001$), suggesting that candidates who are intellectually curious and gregarious tend to receive higher interview scores. This study suggests that clinical performance in dentistry is more closely related to structured interview scores, whereas academic performance in the early years of dental school is more associated with cognitive ability test scores.

Harasym et al. (1996) assessed the inter-interviewer reliability and validity of interviewers' judgements of non-cognitive traits in medical school applicants (using simulated actor-candidates). Interviewers (n = 25; some experienced, some novice) conducted unstructured interviews with both real (n = 155) and simulated candidates (SC, n = 6), although interviewers were blind to which candidates were real vs. simulated. SCs portrayed poor, average and good candidates. Findings showed that interviewers were accurate (correctly rated SC performance) in 56% (20/36) of interviews. Good candidates were identified in 67% (8/12) of interviews, average candidates in 41% (5/12) and poor candidates in 58% (7/12). In the inaccurate ratings, incorrectly rated 'good' SCs were rated as average but never poor. But 'poor' candidates were rated as both average and poor. Average SCs tended to be rated incorrectly as poor. Experienced interviewers were significantly more accurate (correctly rated 70% of SC interviews) than novice interviewers (correctly rated 31% of SC interviews). The study suggested that unstructured interviews lack validity and that it is important to use experienced interviewers.

Hayden et al. (2005) examined potential predictors of performance in an emergency medicine residency programme. Resident (n = 54) application data, including average interview scores, were submitted to regression analyses to predict faculty-rated performance at graduation (overall, clinical, and academic). Interview scores failed to predict success in the programme, although no details were given relating to the format and structure of the interview. Medical school attended and 'distinctive factors' (e.g. being a champion athlete or officer of a national organisation) predicted overall performance, and dean's letter of recommendation predicted overall success in some analyses.

Collins et al. (1995) studied the selection process at a medical school in New Zealand and examined scores on interviews, a group exercise, a school report from the principal, and national exam scores. Candidates (n = 79) were interviewed twice – each time by a panel of two trained interviewers using a structured protocol. Each panel member independently rated the candidate on seven non-cognitive attributes (e.g. communication, maturity). The group exercise and school report also assessed non-cognitive attributes. These three measures of non-cognitive attributes (interview, group exercise, and school report) were fairly highly correlated with each other (ranging from $r = 0.43$ to 0.62, $p<0.0001$), and were not correlated with a national exam score. This provides some evidence that the non-cognitive assessments measure similar attributes (convergent validity) which are not related to an academic test score (divergent validity).

Metro et al. (2005) examined the relationships between application data, including interview scores, and performance after the first and second years of an anaesthesiology residency programme in the USA (n = 18). Few details are given on the format or structure of the interview, but applicants are assessed by residents and faculty (in 4-5 individual interviews). Interviewers rate applicants on several attributes (e.g. personality, intellect), and also rate their credentials (e.g. test scores, letters of reference) to inform the selection committee score. No significant correlations were found between the selection committee score and any of the performance indicators used during the residency, including faculty year-end evaluations and national percentile ranking on the In-Training Examination (ITE). Although the relationship between interview score and performance is not isolated in this study, the selection system of which the interview is a part failed to demonstrate any relationship with performance.

Hall et al. (1992) investigated the nature of the relationship between admission interview scores and dean's letter ratings with students entering a UK medical school (n = 62). Interview scores were based on both academic and non-academic criteria. Results showed a significant correlation between admission interview scores and dean's letter rating ($r_s$= .33, p = .001). The authors recommend the use of interviews as a way to identify applicants most likely to be strong, competitive performers (according to the Dean's letter) in residencies four years later.

Patrick et al. (2001) assessed the reliability and validity of a new structured interview for admission to a US medical school residency programme. In pairs, trained faculty interviewers (n = 73) interviewed 490 applicants to medical school. The predictive validity of the structured interview was tested alongside traditional admissions measures (GPA, Medical College Admissions Test [MCAT] scores, and an in-house evaluation form designed to assess non-cognitive attributes) to predict admission status (accept/reject/wait-list). Using regression analysis, the three traditional measures predicted 16% of the variance in admission status (p<0.001). The structured interview predicted an additional 20% of the variance in admission status (p<0.001), over and above that predicted by traditional measures. However, this study assessed admission status, not actual performance. The authors also examined the relationships between the predictors, but found low or no correlation between the traditional measures and the structured interview, suggesting that they measure different qualities. Interestingly, there were no correlations between interview scores and the in-house evaluation, despite both measuring non-cognitive attributes.

Tran and Blackman (2006) investigated the validity of one on one interviews and group interviews for predicting academic potential (as measured by a cognitive ability test) in 91 undergraduate students in the US. Participants were randomly assigned to take part in a one on one interview (10 minutes) or a group interview (30 minutes). In the group interview, everyone responded to a question from all three interviewers. Results showed that interviewers in the one on one format were significantly better at predicting the applicant's academic potential. In the group interview the order in which the applicants were asked to respond had a significant effect on the quality of their responses. The authors concluded that one on one interview format produces significantly more accurate and valid predictions of a candidates academic potential than would the group interview format.

Coates (2008) carried out a study to assess the criterion validity of Graduate Australian Medical School Admission Test (GAMSAT) to predict performance in medical school (over and above GPA and interview scores). The study looked at the relationship between GAMSAT and concurrent GPA and interview measures, and which combination of GAMSAT, GPA and interview measures best predicted Year 1 performance. Results showed that interviews do not help to predict performance over and above GAMSAT and GPA. There is some evidence for divergent relationships between GAMSAT and GPA/interview scores, suggesting they measure different things, but these relationships were mixed and varied across institutions. The study suggested that in some cases, interviews reduced the predictive validity of the combination of selection methods. Interviews lowered the predictive power of the model for some institutions, but raised it for others. It was difficult to separate these differences without information on the interviews (e.g. whether they were structured or unstructured).

Spina et al. (2000) assessed the relative importance of different selection factors when evaluating an application for entry to oral and maxillofacial surgery (OMFS) residency. Questionnaires were returned by 71 residency programmes (75.5% response rate), and all programmes indicated that they used selection interviews. The personality and appearance of characteristics that are observed in an interview and rated as important include being energetic, confident, honest, organised, and verbally fluent (rated as ‘positive’ by at least 90% of respondents). Being aggressive and anxious were rated as negative factors by 26% and 32% of respondents, respectively. However, the study is that it presents opinions of unspecified respondents in the programme, and no data on whether the factors believed to be important have any validity in predicting actual performance.

Thordarson et al. (2007) compared the initial ranking of incoming residents (n = 46) in an orthopaedic surgical residency programme in the USA to their performance ranking at graduation. Initial ranks were based on an interview and consideration of an applicant’s file. Correlations between initial ranking and ranking at graduation (according to 4 different faculty members) varied from Spearman correlation coefficient = 0.19 (ns) to 0.37 (p = 0.01). However, initial rank was significantly related to 4th-year Orthopaedic-In-Training Examination (OITE) scores (0.34, p = 0.02) and American Board of Surgery (ABOS) Part 1 exam scores (0.36, p = 0.01). This suggests that interview scores, when considered alongside application information, are related to exam performance; although the unique contribution of interviews is unknown. In addition, the study questions the reliability of faculty assessment of resident’s performance.

### *4.4 Discussion: Interview Validity*

Research described in this review indicates that evidence for the validity of interviews is somewhat mixed, with examples of good predictive validity as well as articles failing to find any relationship between interview scores and later performance. This variability may be due to the range of different interview formats and scoring methods used, which have been found to affect validity (Campion et al., 1997). Frequently, details about the interview are sparse, making categorisation (e.g. structured, semi-structured, unstructured) and comparisons difficult, although in general structured interviews typically offer higher validity than unstructured interviews.

Structured competency-based interview (CBI) scores have been found to predict a large amount of incremental variance in residency performance four years later, over and above traditional (presumably unstructured) interviews (Altmaier et al., 1992). Traditional interviews predicted only 2-4% of different performance measures, whereas adding CBI scores increased the explained variance to between 18-26% (depending on the performance measure). In addition, parts of the traditional interview negatively predicted performance (high scores in the interview predicted poor performance), whereas parts of the CBI positively predicted performance (high scores in the interview predicted high performance). Positive correlations have been reported between structured interview scores and clinical performance as a resident, ($r = 0.60$; Olawaiye, 2006) and clinical performance in dental school ($\rho = 0.31$-$0.44$, corrected for unreliability and range restriction; Poole et al., 2007). Personal characteristics assessed at interview have been found to relate positively to clinical performance of surgical residents ($r = 0.21$-$0.46$) and negatively related to whether they are identified as a ‘cause for concern’ ($r = -0.55$), but are not related to exam performance (Brothers & Wetherholt, 2007). However, these latter correlations may be inflated as the same faculty provided ratings both at interview and for performance evaluation. One study also suggested that one-to-one interviews have greater validity than group interviews (Tran & Blackman, 2006).

Other studies have not found relationships between interview scores and performance. Basco (2000) found no significant correlation between interview scores and third year OSCE performance in medical school, although the interview was not structured. Hayden et al. (2005) found that interview scores failed to predict performance in an emergency medicine residency programme, and Metro et al. (2005) found no relationship between a selection score (based on interviews as well as other selection methods) and residency performance, although neither study described the interview format.

Further studies have found mixed results, with interviews predicting some elements of performance, but not others, or demonstrating variability across institutions. Coates (2008) reported that interview scores could either increase or decrease the predictive validity of a selection model using GPA and the Graduate Australian Medical School Admissions Test (GAMSAT), depending on the institution. Heintze et al. (2004) found that interview scores were negatively related to course failure ($r = -0.25$) and positively related to ratings of interpersonal competence ($r = 0.31$) and empathy ($r = 0.30$) in dental school, but that they did not predict these outcomes over and above other selection tools (e.g. spatial ability tests). Interviews were also not related to results on an academic course or to course breaks/drop-outs. However, Hoad-Reddick et al. (1999) found that dental students who scored highly on interview ratings of leadership were less likely to fail the course, but other attributes assessed at interview did not predict course performance. Basco et al. (2008) reported significant but weak correlations ($r = 0.13$-$0.15$) between (presumably non-structured) interview scores and performance on an OSCE in the final year of medical school, but the small effect is of little practical value. In a meta-analytic review of the validity of interviews in healthcare disciplines, Goho & Blackman (2006) reported $r = 0.06$ for the prediction of academic performance, and $r = 0.17$ for the prediction of clinical performance, but the low effect sizes indicate that interviews are of little or no practical value.

The inconsistencies in the results relating to interview validity indicate there is a need for additional research to establish the predictive validity of structured interviews for foundation year performance. Research in other fields indicates that interviews can predict job performance, but they need to be structured as well as carefully developed and conducted (Campion et al., 1997).

In occupational psychology, several meta-analyses have demonstrated the validity of interviews and the importance of structure. Wiesner and Cronshaw (1988) reported a validity of .47, corrected for criterion unreliability and range restriction in the predictor, but they found superior validity for structured interviews (.62 versus .31) when examined separately.

McDaniel et al. (1994) reported an overall corrected validity of .37, but again found higher validity for structured versus unstructured interviews (.44 and .33, respectively). However, Huffcutt and Arthur (1994) found there was a point of diminishing returns in adding structure to interviews. In a meta-analysis, they identified four levels of structure relating to question standardization: Level 1 was wholly unstructured, with an absence of formal constraints; Level 2 was characterised by minimal constraints, typically limited to standardisation of the topics covered in the interview; Level 3 involved the pre-specification of questions, but interviewers were free to choose between alternative questions and probe candidates with follow-up questions; Level 4 was complete standardisation, with identical questions and sequencing, and no probing or deviations. They reported predictive validity estimates for supervisory performance ratings, corrected for unreliability and range restriction, of .20 for Level 1, .35 for Level 2, .56 for Level 3, and .57 for Level 4. It is clear that increasing the standardisation of questions tends to enhance the predictive validity of the interview. However, the trend asymptotes at Level 3, which is suggestive of a ceiling effect for standardisation. This implies that increasing the standardisation beyond Level 3 yields no demonstrable benefit for predictive validity. However, they did detect more variance in validities at higher levels of structure, which may indicate the presence of other potential moderators of this effect. For example, the degree of interviewer training may be important, with highly trained interviewers compensating for less standardisation of questions. Huffcutt and Arthur's results are also important as both interviewers and interviewees tend to rate interviews more favourably that, whilst fair, are not overly rigid (Dipboye, 1994; Hysong & Dipboye, 1999).

### *4.5 Studies investigating user reactions to interviews*

Milne et al. (2001) surveyed 53 current interns (87% response rate) in an internal medicine programme in the USA regarding their perceptions of interviews in the selection process. Interns reported that the most important goals for the faculty interview were: to learn about the programme, to sell themselves, to assess faculty satisfaction with the institution, and to determine their own interest in the programme. A large majority (86%) agreed that the faculty interview was a necessary part of the interview day and most (65%) felt it had impacted their ranking at selection. When asked to rate the acceptability of alternative interview models, the majority felt it would be unacceptable to have either: no faculty interview (93%), one faculty interview at 60 minutes (61%), or two resident interviews at 30 minutes each (59%). The most acceptable model (89%) was two faculty interviews at 30 minutes each, although the majority agreed that it would be acceptable to have two faculty and one resident interview at 20 minutes each (67%) or one faculty and one resident interview at 30 minutes each (62%). This study provides evidence that candidates see the interview (particularly with faculty) as an important part of the selection process and assesses the acceptability of different combinations of interviewers and interview durations from a candidate's perspective.

Thordarson et al. (2007) investigated the perceived importance of different selection criteria according to faculty interviewers (see 'studies investigating the reliability of interviews' section for details). The 4 interviewers generally agreed on the rank order of importance of different selection criteria in their decision-making. Interviews were ranked 4th, after medical school grades and/or Alpha Omega Alpha status, letters of recommendation and rotation evaluations, and United States Medical Licensing Examination (USMLE) Part 1 scores; and before medical school attended, personal statement, research experience, other activities, and gender/ethnicity.

Courneya et al. (2005) compared structured panel interviews with unstructured individual interviews for the selection of medical school students (see 'studies investigating the reliability of interviews' section for details). Following the interviews, candidates reported that the structured panel interviews put them at ease, allowed them to express themselves and enabled them to leave an accurate impression of themselves, compared to the unstructured interviews. Interviewers in the panel (including clinicians, academics, and members of the community) reported that they were able to create a comfortable climate for the candidate, there were sufficient numbers of interviewers, and that the diversity in the panel was beneficial. No comparable results were available from interviewers in the unstructured group.

Spafford and Beal (1999) surveyed 109 applicants (69.4% response rate) to an optometry programme at a Canadian hospital following an admissions interview, and compared male and female responses. They asked applicants to rate the purpose(s) of an ideal interview and their perceptions of the actual purpose(s) of the interview for the optometry programme. There were no significant group differences between men and women's expectations of the ideal interview or between their perceptions of the programme interview. Both men and women reported that the programme interview was less geared towards addressing public relations and verifying information than an ideal interview should be. Furthermore, women felt that the purpose of the programme interview was less focused on recruitment and gathering information than an ideal interview should be. This study indicated that there is a disparity between actual and ideal interview purpose, as perceived by applicants, with a greater disparity for female applicants.

Tran and Blackman (2006) investigated the validity of, and user reactions to, one on one interviews and group interviews for predicting academic potential (see section on studies investigating validity for details). Students (n = 78) were asked to report their perceptions of fairness. Interviewees reported negative perceptions of the fairness and appropriateness of the group interview method. However, the study is not based on medical selection interviews.

Chapman et al. (2003) compared interviewee reactions to three types of interview: face-to-face, telephone, and video. Interviewees (n = 802, 86% response rate) were Canadian university students using the campus recruitment centre, which was used by employers to fill positions. Interviewees did not know what type of interview they would receive, minimising self-selection to a particular type of interview. Post-interview questionnaires indicated that face-to-face interviews were perceived as fairer than telephone ($t(694) = 8.44$, $p<0.01$, $d = 0.83$) and videoconference ($t(648) = 6.65$, $p<0.01$, $d = 0.77$) interviews, but there was no difference between telephone and videoconference interviews. Also, after controlling for pre-interview intentions, face-to-face interviews resulted in significantly higher post-interview job acceptance intentions than telephone interviews ($t(706) = 2.02$, $p = 0.04$, $d = 15$). These results suggest that, although there are potential cost-savings of using technology-mediated interviews, applicants may have more negative reactions to them when compared to face-to-face interviews.

Westwood et al. (2008) explored candidate reactions to a selection process, including a structured interview, for selection of cardiology specialty trainees at the London Deanery. Interviews involved three panels and candidates were asked the same initial stem questions, which were followed up with 'probe' questions. Candidates (n = 94, 80% response rate) were sent a questionnaire after the interview but before selection results were announced. In general, candidates reported favourable reactions to the structured interview: they were satisfied with the process and felt it was objective, believed the panel composition and interview duration were appropriate, were asked relevant questions, and could express their individuality. Candidates also reported that it would not be helpful to require a short presentation (either on a predetermined or self-selected topic) as part of the interview. These results indicate that candidate reactions to a structured interview were generally positive.

### *4.6 Panel composition*

Milne et al. (2001) investigated the effect of replacing faculty members with residents in the selection process for a US internal medicine residency programme. Specifically, they compared faculty and resident ratings of paper applications and interviews, and the effect of varying the composition of the interview panel (2 faculty, n = 93; 1 faculty + 1 resident, n = 89; 2 faculty + 1 resident, n = 85), compared to the 'control' formal interview process (two separate interviews with faculty). Scores assigned by both faculty and residents correlated highly with the formal interview score (r = 0.72 and 0.65, respectively). Varying the panel composition had no significant effect on applicant rank, but results suggested that residents were not as good as faculty at evaluating paper applications and they were consistently more lenient than faculty.

### *4.7 Influence of interviews*

Nowacek et al. (1996) aimed to determine whether medical school admission interviewers change their evaluations and impressions of applicants as a direct result of the interview. Applicants to a US medical school (n = 419) were interviewed by members of the admission committee in two 30-minute sessions. Before the interview, interviewers reviewed and rated the applicant's folder. After the interview the applicant was rated again. Results showed that most of the applicant's ratings were significantly changed by the interview (change of up to 0.34 on a 5-point scale, p<0.01). Stepwise multiple-regression showed that the communication and interpersonal skills ratings contributed most to predicting the overall impression both before and after the interview. There were significant differences in before and after-interview ratings on Overall Impression for those accepted for admission, put on the alternate list and rejected. The Overall Impression rating of the accepted group increased an average of .21 points (p<0.01) and the alternate-list groups by .06 (ns), whilst the rating of the rejected group decreased by .25 (p<0.01). The changes (albeit not large) between pre and post-interview ratings of medical school applicants suggest that the interview does influence the selection of individuals for medical school.

Elam et al. (2002) explored the relationship between committee members' background characteristics and their propensities to change votes in a context that allowed for deliberation about applicants to medical school. Committee members (n = 18) at a US medical school varied in their backgrounds (faculty, student, administrator/community member) and years of committee service (<5 years or >5 years). Results indicated that negative vote changes occurred equally among the 3 categories of background ($F_{(2.25)}$) = .03, p = .97). However students were more likely than faculty members and administrators/community members to change in a positive direction, and faculty members were more likely than administrative/community members to change in a positive direction ($F_{(2.25)}$) = 20.08, p = .0001). More experienced committee members with 5 years or more of service, were also significantly more likely to change votes in a positive direction (mean=8.5%) than were less experienced committee members (mean=5.3%) (F($_{1.25}$) = 11.22, p = .002), but equally likely to change votes in a negative direction (experience high mean = 10.5%, experience low mean=9.4% ($_{F(1.25}$) = 0.43, p = .52

Galazka et al. (1994) surveyed residency directors of U. S. family practice programmes (N = 282, 78% response rate) on how they recruit and select residents. Results indicated that the most important factor in the selection of candidates was interviews (51%), followed by candidates' performances on clinical rotations (36%). The Dean's letter, personal letters, medical school transcripts and reference letters were all rated most important in selecting first year residents, but by less than 8% of the directors. In conclusion, residency directors' selection decisions were highly influenced by candidates' performances in interviews and on clinical rotations.

## *4.8 Introduction of bias in interviews*

### *4.8.1 Gender*

Frantsve et al. (2003) examined the effects of gender and personality traits on selection for an oral maxillofacial surgery residency programme in the US (n = 47). Personality was assessed by a standardised self-report personality measure (Adjective Check List, ACL) and by interviewers on five personality traits (e.g. assertiveness, friendliness) in 3 x 20-30-minute interviews. There were no significant correlations between the ACL and applicant ranking, but all of the interviewer's trait ratings were significantly related to applicant rank, ranging from r = -0.33 for stress tolerance and for confidence, to -0.52 for assertiveness (all p<0.05). Successfully matched applicants were rated by faculty as being more assertive, and were more likely to elicit sympathy and emotional support from others according to the ACL. A gender difference was detected in interviewers' ratings: female applicants were perceived to be less able to cope with stress effectively than male applicants (F(1,44) = 9.49, p = 0.004), despite there being no evidence of stress management difficulties in female applicants compared to the general population. This study suggests that interview scores may be somewhat affected by gender, although this may be specific to the specialty.

### *4.8.2 Prior knowledge of academic scores*

Shaw et al. (1995) assessed the relative contribution of Medical College Admissions Test (MCAT) and Grade Point Average (GPA) scores in predicting unstructured interview scores when interviewers were aware versus unaware of test scores. The effects of race and gender were also investigated. Interviewers rated 226 US medical school applicants with access to their MCAT and GPA scores, and a further 245 applicants without access to their MCAT and GPA scores. Regression results showed that GPA was the strongest predictor of non-cognitive trait ratings (NTR) in both cases, although it accounted for considerably less variance in NTR when interviewers did not have access to academic scores (15.7% vs.7.3%). Male and female interviewers rated female applicants higher than male applicants on NTR, and there was evidence of a race bias in favour of African-Americans (AA), such that AA applicants were rated higher on NTR than non-AA applicants, possibly as a result of affirmative action policies at the university. The study suggests that interviewers should be blind to academic information.

Smilen et al. (2001) investigated the influence of advance knowledge of board scores (United States Medical Licensing Examination, USMLE, Part 1 scores) on interviewers' assessments of residency applicants (n = 152) in a US obstetrics and gynaecology programme. Results found no difference in overall mean interview scores when interviewers had access to USMLE scores and when they did not. When interviewers had access to USMLE scores, they were significantly correlated with interview scores (r =0.64, p<.0001), but there was a slightly negative correlation between USMLE and interview scores (r = -

0.06) when interviewers were unaware of them. This bias was evident among all levels of seniority of examiners, although it was more apparent among the junior members of the faculty. These results suggest that the availability of markers of academic achievement (e.g. USMLE board scores) to interviewers may bias the evaluation of the interview, and may increase the likelihood of a 'halo effect'.

Miles et al. (2001) investigated the effects of blinding interviewers to application file data for 132 candidates for a surgical residency programmes at two sites in the US. Each candidate was interviewed twice – one interviewer was blind to application data and the other had access to the file. Results showed a significant difference between mean blinded and unblinded interview ratings at one site (mean ± SD:23.0 ± 17.7 versus 32.6 ± 23.1), but not at the other (17.9 ± 20.5 versus 14.3 ± 9.1). Interview ratings correlated significantly with USMLE scores at one site, but in opposite directions for blinded (r=0.32, p = 0.003) versus unblended interviews (r = -0.32, p = 0.003). These correlations were not found to be the same at site 2. These results suggest that assessment of residency candidates by surgical faculty may be influenced significantly by the provision of application data to interviewers before the interview, although there are differences across programmes.

### *4.9 Discussion: Implementation and User Reaction*

Interviews have generally been perceived positively by both interviewees and interviewers. The main perception reported by interviewers and interviewees was that interviews were an important and necessary part of the selection process. Courneya et al. (2005) compared structured with unstructured interviews, and reported that structured interviews put interviewees at ease and allowed them to express themselves more than unstructured interviews. These findings were also reflected in the interviewer perceptions. Westwood et al. (2008) only looked at structured interviews and found that they were perceived positively and objectively. In addition, two studies examined perceptions of fairness and interview type: face-to-face interviewers were seen as fairer than telephone and videoconference interviews, and individual interviews are seen as fairer than group interviews. However they are more costly for both interviewees and the employing organisation.

Several studies discussed the introduction of bias in the interview method. However it was often unclear whether the studies looked at structured or semi structured interviews and they were often specific to one site or one specialty. One study reported that interview scores could be affected by gender bias. Another bias that could be introduced into interviews was prior knowledge of academic scores which could create a 'halo effect' on interview scores. It was also found that interviews influenced decisions and rankings made in the selection process.

### *4.10 Conclusion*

Interviews can be moderately reliable and valid predictors of performance, but there is considerable variation in reported results. Structured interviews are more reliable than unstructured interviews. Validity evidence is mixed, ranging from studies finding no relationship (or even a negative relationship) between interview scores and performance, to studies demonstrating good incremental predictive validity. This variation may be due to methodological differences in the studies: the term 'interview' describes a range of different formats, which are often not described, but that may affect both reliability and validity. Non-medical studies provide evidence in support of structured interviews, and indicate that adding structure improves predictive validity, although complete standardisation of the interview process adds little additional predictive validity and may be disliked by users. The increased reliability and validity associated with structured interviews also offer a greater degree of legal defensibility than unstructured interviews.

User reactions to interviews were generally positive. Interviews were regarded as a necessary part of the selection process and they were given considerable weight by programme directors. Candidates have reported that structured interviews put them at ease and allowed them to express themselves more than unstructured interviews, and interviewers have also responded positively to structured interviews, indicating that they were able to create a comfortable climate for the candidate. Candidates have also reported that they regard structured interviews as objective assessments. However, interviews are a costly and resource-intensive selection procedure, both for the employing institution and the candidates.

Gender bias may affect interview scores, and prior knowledge of academic credentials may create a ‘halo’ effect that influences interview scores. Unstructured interviews are typically more prone to biases.

There is a need for additional research on interviews with respect to their reliability and validity for the Foundation Programme, including across institutions. If used, interviews should be structured, conducted by trained interviewers, and based on a thorough job analysis to identify key competencies required for successful performance as a doctor.

## 5. Multiple Mini Interviews (MMIs)

Multiple mini interviews (MMIs) were first developed by Kevin Eva and his colleagues in Canada in an attempt to address context specificity problems and biases associated with traditional interviews. They argued that a candidate's performance in a single interview may not demonstrate their true abilities and may be particularly susceptible to biases (e.g. similarity of background between interviewer and candidate), but a more accurate assessment is possible if multiple samples of performance across contexts are obtained.

Eva et al. (2004a) described MMIs as a 'multiple sample approach to the personal interview' (p.314) and demonstrated how MMIs could be used as a reliable, flexible and feasible admissions tool. Based on an objective structured clinical examination (OSCE)-like model, MMIs are composed of a series of stations, each with its own interviewer, and candidates are rated on salient competencies or qualities at each station. However, unlike an OSCE, MMIs are typically subjective and non-clinical. The use of multiple stations minimises the problem of context specificity and the use of multiple interviewers dilutes the impact of interviewer effects (e.g. leniency, similarity of background between interviewer and candidate) on a candidate's score. Stations can be adapted to assess a range of skills and competencies, which may be cognitive (e.g. critical thinking) and/or non-cognitive (e.g. empathy), and may involve the interviewer acting as an observer or asking questions and interacting with the candidate. For example, to assess communication skills, the interviewer may observe the candidate talking with a simulated 'patient'; and to assess ethical decision-making, the candidate may be asked to discuss issues relating to a doctor's recommendation of homeopathic remedies to his patients, despite a lack of evidence for their effectiveness. Stations can be designed such that the tasks/questions assume no clinical knowledge, and some stations can also include standard interview questions (e.g. Why do you think you will be a good doctor?). Given the range of different tasks and questions that can be used to assess a competency in MMI, Eva et al. (2004a) suggest that candidates will be less able to rehearse responses than for a traditional interview where they may be asked a well-known question (e.g. tell me about a time that you solved a difficult problem). Critically, the MMI typically requires candidates to actually demonstrate a competency, rather when describe an occasion that they demonstrated a competency.

### *5.1 Studies investigating the reliability of MMIs*

Eva et al. (2004a) discussed the biases and variable levels of reliability reported in studies on traditional admissions interviews, and described MMIs as a new selection method aimed at minimising such biases and the problem of context specificity. They conducted two studies: a pilot to test the MMI with 18 graduate students at McMaster medical school in Canada, and a larger-scale study (n=117) to evaluate the use of MMIs in the selection of medical students at the same school. In the pilot study, candidates rotated through six MMI stations, each lasting 8 minutes, and were rated by two interviewers at each station. In the main study, medical school applicants rotated through ten MMI stations, each lasting 8 minutes, with one interviewer per station. In both studies, candidates were rated on their communication skills, strength of arguments, suitability for the health sciences, and overall performance using a standardised evaluation form with 7-point scales. The overall test generalisability (i.e. G-coefficient the reliability of the average of the 12 ratings, one from each interviewer) was 0.81 for the pilot, and 0.65 for the main study. In the main study, the variance attributable to the candidate-station interaction was five times greater than that assigned to the candidates themselves. The MMI score did not correlate highly with any other admissions tools used at McMaster (personal interview, r=0.19; simulated tutorial, r=0.32; undergraduate grade, r=-0.23 and autobiographical sketch, r=0.17). Despite these low correlations, those who were admitted to the medical school received significantly higher scores on the MMI (mean=5.30/7) than those who were not (mean=4.83/7), even though MMI scores were not available to the admissions committee. In addition, post-MMI surveys provided positive feedback on the process from both interviewers and candidates, and the authors suggest that MMIs may be more cost-effective than traditional interviews.

Goodyear et al. (2007) assessed the reliability of MMIs for the selection of candidates for paediatric Senior House Officer (SHO) posts in the West Midlands Deanery (UK speciality training). Candidates (n=123) rotated through three MMIs (5 minutes each) and were rated independently by two trained and experienced interviewers at each station. MMI stations were question-based, rather than observation-based, and candidates were asked the same three standardised questions which explored competencies validated for paediatric training (insight and reflection, communication skills, clinical knowledge). Inter-interviewer reliability was good across all six stations. (Cronbach's alpha=.74 - .84). There was good correlation between stations (G coefficient=.80 for 3 stations and 2 raters). Using Decision studies and generalisability theory, the authors theoretically varied the number of interviewers per station (from 1 to 3), and the number of stations in the MMI (from 3 to 8). G-coefficients ranged from 0.71 (3 MMIs and 1 rater) to 0.93 (8 MMIs and 3 raters).

Donnon and Paolucci (2008) developed three medical judgement vignettes, similar in design to MMIs, in which applicants discussed scenarios and were asked questions relating to ethical dilemmas (moral), relationships with patients and their families (altruistic), and collaboration and clarification with staff and colleagues (dutifulness). First year medical students (n=29) at the University of Calgary in Canada were scored independently by two interviewers at each station using structured/anchored rubric based on Kohlberg's theory of moral development, which assessed the logic of their reasoning. The mean inter-rater reliability coefficient between the two independent judges was Kappa=0.95 across the three stations. Correlations between scores on pairs of vignettes ranged from r = 0.22 to 0.49. The generalisability coefficient for 2 judges and 3 stations was 0.70. Using generalisability theory, the authors predicted the effects of varying the number of raters and stations. They found that increasing the number of raters from 2 to 3 did not improve reliability, but increasing the number of stations did: with 4 stations and 2 raters the generalisability coefficient would be 0.75, and with 6 stations and only 1 rater the generalisability coefficient would be 0.81. The authors concluded that the medical judgement vignettes offered a reliable method to assess non-cognitive attributes, and that reliability could be enhanced by including more stations, but not by using more raters (possibly because the structured interview process and scoring rubric resulted in high inter-rater reliability already). However, the study was somewhat limited by its small sample size and use of volunteer participants who had already been selected into medical school.

Lemay et al. (2007) investigated whether MMIs could reliably measure and distinguish non-cognitive attributes, and differentiate between accepted and wait-listed candidates. Applicants to the University of Calgary Medical School in Canada (n = 281) rotated through ten 8 minute stations and were rated by one interviewer per station. Nine stations required candidates to discuss scenarios, and each was designed to measure a distinct non-cognitive attribute. At each station, candidates were rated on 5 criteria using a 10-pt scale. Reliability analyses indicated that internal consistency reliability for each station was high (Cronbach's alpha=.97 to .98), indicating high item cohesiveness among the subscales at each station. To assess whether stations measured single or multiple constructs, the correlations between total scores at different stations were examined, and an exploratory factor analysis (EFA) was conducted on the data. Correlations between different station scores ranged from .04 to

.36, most were low but there were significant positive correlations. However, EFA resulted in a 10-factor solution (each station formed one factor) accounting for 91.9% of the variance. This suggests that the MMI measured multiple distinct constructs, although the data cannot demonstrate that the intended attributes were captured at each station. The results also suggested that MMIs could differentiate between accepted and wait-listed candidates, as there were significant differences in mean scores between accepted and wait-listed candidates for each station. No significant differences were detected between accepted and wait-listed candidates with respect to gender or age, suggesting that the MMIs were fair and would be legally defensible (although more detailed analyses would be required to establish this).

Roberts et al. (2008) conducted a prospective study to establish whether interviewers can make reliable decisions about applicants when selecting candidates for entry to a graduate-entry medical programme using a pre-professionalism framework and the MMI format. The sample consisted of applicants for a graduate entry PBL programme in Sydney, Australia (n = 485 applicants) and interviewers (n = 155). Data was from a high stakes admissions process and content validity was assured through using a framework based on international criteria for sampling expected behaviours of entry-level students. Results showed that the reliability for an eight question MMI was 0.7. To achieve reliability scores of 0.8 a 14 question MMI would be required. For one MMI question and one interviewer 22% of the variance between scores reflected candidate to candidate variation. The remaining 78% would reflect unwanted factors for example, interviewer stringency and interviewer question taste. Interviewer stringency and question taste could not be separated out. MMIs were found to be a moderately reliable form of assessment. A high source of error relates to aspects of interviewer subjectivity which suggests that interviewer training would be beneficial as the interviewers only received a one hour workshop and written material. Findings also suggest that higher numbers of questions asked in the MMI are important for enhancing reliability.

A study by Ziv et al. (2008) aimed to devise a simulation-based assessment centre for the evaluation of personal and interpersonal qualities of medical graduates. The study was carried out at the Sackler School of Medicine at Tel Aviv University, Israel. Participants included 283 applicants to the medical school in 2004 and 280 applicants in 2005. A survey was conducted among senior medial professionals to identify the personality and behavioural characteristics required of medical students and doctors. Thirty professional actors were used to role-play as simulated patients. Workshops were run to standardise scenarios and 150 doctors were recruited as raters for the tests which were carried out over three days. Each candidate had two six hour sessions with 96 students being tested each day. Candidates' qualities were scored by faculty and the simulated patients using a structured assessment focused on the personal qualities listed above. Results showed that inter-rater reliability was moderate (0.62-0.77). Test-retest (after one year) was found to be moderate to high (0.7). There was no correlation between MMI stations and the cognitive test score (r = 0.02), indicating that they are measuring different qualities, and providing some evidence for the discriminant validity of the MMI. In addition, MMI stations, which were purported to measure non-cognitive attributes, were significantly correlated with other indicators of non-cognitive attributes (r = 0.25 to 0.38). The 'Selection for Medicine' (MOR) tool seems to be a reliable tool for measuring non-cognitive attributes in medical school candidates, showing high content and face validity. The limitations of this study are that the cost effectiveness of the MOR tool is a weakness, the study was conducted at a single site and that there are no studies as yet on the prediction of future performance.

Eva et al. (2004c) carried out a study to assess the consistency of ratings assigned by health sciences faculty members relative to community members during an MMI. The population consisted of 198 candidates to an undergraduate MD programme in 2003 at McMaster University, Canada who were invited to take part in an MMI for research purposes (n=54 participants); 36 examiners (n =18 health sciences faculty members, n = 18 community members). A nine station MMI was created; three stations had two faculty examiners, three had two examiners from community; three had one examiner from each group. Raters completed a 4-item evaluation form for each candidate. Generalisability theory was used to examine consistency of ratings provided within each of the three subgroups. A decision study was used to determine the optimal combination of stations and raters. At the end of the MMI all candidates and examiners were asked to complete questionnaires regarding their perceptions of the process. Results showed the overall test reliability was .78 (the reliability of the average of all 18 ratings). Generalisability was highest for the three stations staffed by community members (.58), second highest for those staffed by two faculty members (.46) and lowest for pairings of community member and faculty member (.31). Each pairwise difference was statistically significant. The Decision study showed that increasing the number of stations has a greater impact on the reliability of the interview than increasing the number of raters within each interview. Candidates and examiners responded positively to the MMI, for example, with regard to the adequacy and clarity of the instructions prior to the MMI and each station and accuracy of portrayal of abilities by/of candidates. The number of stations is a more important determinant of the overall reliability of the MMI than the number of panellists in the interviews. The limitations of the study were that volunteers and the participants were aware that the results would have no impact on admission process and therefore their motivation may not be as high. This paper indicates that the MMI can be a reliable evaluation instrument for medical school admissions and highlights the contribution of the number of stations to overall reliability. Applicants vary from one interview to another, so scores from one interview will be a poor predictor of performance in a second interview. The study presents findings on the optimal balance between interviews and interviewers.

### *5.2 Summary of reliability of MMIs*

Reliability studies suggest that the MMI is a reliable measure of non-cognitive attributes (e.g. communication skills, empathy), and that reliability is enhanced with the inclusion of more stations, but not necessarily the inclusion of more raters at each station. Reported generalisability coefficients for MMIs included: 0.81 and 0.65 for a 6 x 8-minute station MMI with 1 rater (Eva et al., 2004a); 0.80 for a 3 x 5-minute station MMI with 2 raters (Goodyear et al., 2007); 0.7 for an 8 x 7-minute station MMI with 1 rater (Roberts et al., 2008); 0.70 for a 3 station MMI (total time=15-20 minutes) with 2 raters (Donnon & Paolucci, 2008); and 0.78 for a 9 x 8-minute station MMI with 2 raters (Eva et al., 2004c).

Several researchers also used generalisability theory to theoretically predict MMI reliability if the number of raters or stations was varied. For example, using this approach, generalisability coefficients have ranged from 0.71 for a 3 x 5-minute station MMI with 1 rater to 0.93 for an 8 x 5-minute station MMI with 3 raters (Goodyear et al., 2007). Eva et al. (2004c) found that reliability dropped to 0.45 for a 1 x 8-minute station MMI with 18 raters, and rose to 0.81 for an 18 x 8-minute station with 1 rater. Donnon and Paolucci (2008) reported generalisability coefficients from 0.44 for 1 station and 5 raters up to 0.81 for 6 stations and 1 rater. These studies have demonstrated that increasing the number of stations has a greater impact on the reliability of the interview than increasing the number of raters within each interview (Donnon & Paolucci, 2008; Eva et al., 2004c; Goodyear et al., 2007). Other studies have found moderate inter-rater reliability (0.62-0.77) on role-play stations with simulated patients (Ziv et al., 2008), and higher inter-rater reliability

(Cronbach's alpha=0.74-0.84) on question-based MMIs (Goodyear et al., 2008). Ziv et al. (2008) also reported moderate/good test-retest reliability when the MMI was repeated after a one year interval (r = 0.7).

### *5.3 Studies investigating the validity of MMIs*

Several articles report correlations between MMI scores and performance outcomes, but few report the results of regression analyses, etc, that may be used to assess the relative contribution of selection methods to the prediction of performance outcomes. High correlations between MMI scores and performance demonstrate there is a relationship between the two, but do not control for the effect of other variables. If decisions were based on correlational evidence, then two selection criteria may be adopted that both correlate highly with performance, but one may not predict any incremental variance over the other.

Eva et al. (2004b) made the first attempt to assess the validity of the MMI by examining the relationship between pre-clerkship performance (consists of OSCEs and a multiple choice exam testing medical knowledge – the personal progress inventory, or PPI), the MMI and the traditional admissions tools used by the undergraduate medical programme at McMaster University, Canada. The performance of forty five applicants who had completed the MMI as part of their application to join the medical school at McMaster University and their performance on the traditional protocol was compared to the performance on pre-clerkship evaluation exercises. The results showed that the MMI was the best (and only statistically significant) predictor on the objective structured clinical examination (OSCE), with a standardised coefficient (β) of 0.44 (p<.01). Grade Point Average (GPA) was found to be the best predictor on multiple choice exams measuring medical knowledge (β = 0.54, p<.05), followed by an autobiographical submission (β = 0.45, p<.05). The study shows that MMIs appear to be the best predictor of pre-clerkship performance as measured by OSCEs and supports the hypothesis that the MMI provides a more valid indication of candidates' non-cognitive characteristics than do the more traditional admission tools. The traditional personal interview did not correlate with performance and replicates the work of Basco et al. (2000). The finding that the GPA was the best predictor on medical knowledge tests replicates the findings of Kulatunga-Moruzi and Norman (2002a,b). Limitations of this study are that it is a single site study with a relatively small sample size. However the study highlights how the MMI was the best predictor of OSCEs which are important for measuring clinical practice.

Reiter et al. (2007) assessed the predictive validity of MMIs, as well as undergraduate grade point average (uGPAs) on clerkship performance and scores on Part 1 of the Medical Council of Canada Qualifying Exam (MCCQE). Questions explored in the research were; does the MMI predict clinical clerkship performance? Does the MMI predict national licensing exam performance? How does the predictive validity of the MMI compare with that of more traditional admissions measures of professional qualities? And how does predictive validity of the MMI compare with that of the uGPA? Forty-two participants who had been admitted to medical school at McMaster University in Canada took part in the study. The study evaluated a ten station MMI on a 1-7 point scale they also used the same scale to measure the other standard admission processes (i.e. uGPA, personal interview, autobiographical submission and simulated tutorial). The study also looked at the in–course measures: OSCEs personal progress inventory, clerkship director ratings and encounter cards and licensing exam measures: CLEO (legal and ethical), PHELO (public health), 5 specialties which all have multiple choice questions (MCQ) and clinical decision making (CDM) short answer papers. Findings showed that MMIs had the highest correlation with OSCEs, clerkship director ratings and clerkship encounter cards. Findings also showed that MMI score was the only significant predictor of OSCEs, and MMIs are the best predictor of CLEO and PHELO performances and not of clinical areas. They are also the best predictor of CDMs, whereas findings show that the uGPA is the best predictor of MCQs. The MMI ranked two students (who later failed) lower than other admission measures. The study shows that MMIs predict clerkship performance and professional components of the licensing exam. Prediction is favourable to traditional admission measures, and comparable but complementary to GPAs with MMIs predicting clinical performance measures better. Limitations of the study were the low sample size and that MMIs are not used in the selection process and therefore not deemed as high-stakes. Authors note the restriction of the range in the independent variables. The MMIs may be useful as a selection process into F1 as they are a good way of predicting clerkship and OSCEs.

Bindal et al. (2007) followed up on Goodyear et al. 's (2007) study (see the 'studies investigating the reliability of MMIs' section) and reported correlations between MMI scores (from 3 x 5-minute stations, each with 2 raters) and job performance of paediatric Senior House Officers (SHOs) in the West Midlands Deanery. Senior House Officer performance was rated by educational supervisors (61% response rate) and included questions on communication skills, clinical and personal skills, and career plans. There was a fairly high correlation between MMI scores and communication skills three months into the job (r=.42, p=.007); suggesting that MMIs may be able to predict this element of job performance. However, the relationship between MMI scores and clinical and personal skills (r=.14), and career planning (r=.15) were considerably weaker.

Ziv et al. (2008) provided some evidence for the discriminant validity of their simulation-based MMI stations, which were purported to measure non-cognitive attributes (see the 'studies investigating the reliability of MMIs' section for details). No correlation was detected between MMI stations and a cognitive score (r = 0.02) but significant correlations between MMIs and other indicators of non-cognitive attributes were found (r = 0.25 to 0.38).

### *5.4 Summary of validity of MMIs*

Two studies have investigated the predictive validity of MMIs, using medical school students or applicants to McMaster University in Canada. Eva et al. (2004b) found that MMIs were a better predictor of performance on OSCEs (i.e. clinical practice) than traditional admissions measures (GPA and traditional interviews), whereas GPA and an autobiographical submission were better predictors of performance on medical knowledge exams. Similarly, Reiter et al. (2007) reported that MMIs predicted OSCE performance, clerkship performance (as rated by director ratings and encounter cards), and the professional/non-clinical components of a licensing exam, whereas GPA predicted performance on medical knowledge exams.

Other studies reported correlations between MMI scores and performance outcomes, but they do not control for the effect of other potential predictors. For example, Bindal et al. (2007) found a fairly high correlation of r = 0.42 between MMIs and communication skills (as rated by educational supervisors of paediatric SHOs, after 3 months), but reported low correlations between MMIs and other aspects of job performance (with clinical and personal skills, r = 0.14; with career-planning, r = 0.15). More evidence is required to support the predictive validity of MMIs, particularly studies that use regression analyses to establish the relative predictive power of MMIs and other selection criteria for salient aspects of job performance (e.g. non-cognitive attributes).

Ziv et al. (2008) provided some evidence for the discriminant validity of their simulation-based MMI stations, which were purported to measure non-cognitive attributes. They found no correlation between MMI stations and a cognitive score (r = 0.02) but significant correlations between MMIs and other indicators of non-cognitive attributes (r = 0.25 to 0.38).

### *5.5 Studies investigating the implementation of MMIs*

Dodson et al. (2009) studied the affect of interview duration on MMI outcome and reliability, and tested whether MMI stations could be shortened without compromising reliability or changing applicant rankings. Applicants to Deakin Medical School in Australia (n=175) were assessed at ten 8 minute MMI stations, each one assessing one of 10 core areas: communication skills, professionalism, social justice, evidence use, self-directed learning, teamwork, effective use of resources, career motivation, health promotion, and rural awareness. Half of the stations were conducted under 'control' conditions (8 minute MMI, with ratings made at the end of the 8 minutes) and half were conducted under 'experimental' conditions (8 minute MMI, with ratings made after 5 minutes following a visual signal visible only to the interviewer and again after 8 minutes). Reliability was not greatly affected by the reduction in MMI duration: generalisability coefficients were 0.78 for 8 minute stations and 0.75 for 5 minute stations. Furthermore, there was little difference in rankings based on 5 minute or 8 minute scores, which were highly correlated (0.92 for cumulative scores). Rank did not change for one third of the candidates, and changed by 1-3 positions for the remaining candidates, with the most pronounced changes occurring for candidates with the highest and lowest rankings (where subtle rank changes are unlikely to affect the outcome). Although there was little effect on ranks, mean scores at 5 minutes (M=3.50 out of 5, across all stations) were significantly lower than mean scores at 8 minutes (M=3.70, across all stations). This was also reflected in mean cumulative scores (totals of all stations), which were significantly lower after 5 minutes (17.50) than after 8 minutes (18.50). However, there was no difference between scores at 5 minutes and 8 minutes on 72.5% of the experimental stations (634/875). The authors concluded that reducing MMI duration to 5 minute stations (from 8 minutes) can minimise demands on resources (e.g. interviewer time) without compromising reliability or significantly affecting applicant ranking. However, significant mean differences in scores were found, with lower scores at 5 minutes compared to after 8 minutes. Conclusions should be treated with caution as a major limitation of the study was that the same interviewer provided ratings of the candidate at 5 minutes and at 8 minutes, which may act to reduce differences between scores. Yet even with the same interviewer, scores at 5 minutes were significantly lower than after 8 minutes. If different interviewers were used (as would typically be the case), the difference in scores may be greater, which could have a stronger effect on ranking.

Reiter et al. (2005) investigated whether security violations (i.e. prior knowledge of questions) resulted in undesirable enhancement of MMI performance ratings. The three part comparative study was conducted at McMaster University in Canada. The first study was conducted as a pilot study involving 57 applicants to medical school who had volunteered to participate in the study. Half of the participants were given the MMI stem questions two weeks prior to the interview. The second study involved 384 participants each half received one of two (out of the twelve stem questions) pilot stem questions in advance. The third study involved 38 dual applicants to Occupational Therapy (OT) and Physiotherapy (PT) who experienced the same MMI station on the same day. Findings from study one showed that there was no statistical difference between those who had received the stem questions in advance (mean score 4.97 (SO = 0.46) with those who had not, mean score 4.91 (SO = 0.67). Findings from study two showed that when comparing performance on the station participants received in advance with performance on the other eleven stations revealed no effect of prior exposure. Exposed station mean 4.92 (SO = 1.36) mean performance on stations were not exposed to prior MMI was 4.94 (SO = 0.65) there were no significant differences found between performance in the second exposed stations compared to the ten that were not (mean 4.95, (SO = 0.66)). The third study found that there were no significant differences between the first interview score, mean 3.46 (SO = 0.43) and the second interview score which was mean 3.45 (SO = 0.44). The findings suggest that prior knowledge (i.e. violations) of MMI stations do not influence applicant performance ratings significantly. A possible limitation of this study was that in the first study participants were aware that the MMI scores did not count toward their actual admission decision and therefore those who knew the stem question may not have been as motivated. The study highlights that MMIs are not affected by security breeches and that there is not one formulaic answer to the question.

Griffin et al. (2008) assessed the effects of coaching and repeated testing on the MMI, as well as on an admissions test. Applicants (n = 287, 84% of all students selected for interview) for medical school at the University of Western Sydney completed a self-report survey on their perceptions of the usefulness of coaching, attending other interviews, and having a ‘practice run’ of an MMI. Their MMI scores were also obtained. Half (51%) of interviewees had attended coaching, but the results indicated that coaching did not improve MMI scores, even after controlling for age, sex, and university admissions index (coached M = 3.54, non-coached M = 3.56). However, the coached group had significantly lower scores (M = 3.81, out of 5) on one MMI station (assessing communication skills) than the non-coached group (M = 4.01). There were no significant differences in MMI scores between candidates who had attended other medical school interviews and those for whom the MMI was their first interview (although at the time, no other Australian medical schools were using MMIs). In addition, candidates (n = 17) who repeated the MMI in 2007 (having been unsuccessful in 2006) did not improve their scores on stations that involved new content, but they did make small improvements on stations using the same or similar content. Survey results indicated that candidates believed a ‘practice run’ would be the most effective way to prepare for an MMI, followed by attending other interviews. Coaching was regarded as the least effective tool to prepare for MMIs. In summary, coaching did not improve MMI scores but practice on similar MMI tasks can improve scores.

Rosenfeld et al. (2008) carried out a comparative study looking at the costs of MMIs with traditional panel based interviews at McMaster University, Canada. The study compared the cost of a traditional panel based interview (panel of 3 interviewers) with a twelve station (8 minute) MMI. They compared the cost of the initial setting up and creation of the MMI stations with the interview time, the assessor’s time, staff (planning/implementation/data entry) time and salary, miscellaneous expenses (subsistence, parking) quantity and infrastructure availability (rooms required). The findings showed that MMIs require a greater effort initially in the preparation of the stations and in developing a blueprint specific to the institution. There are a greater number of rooms required to carry out the MMIs relative to those required to conduct panel based interviews. However the panel based interviews cost more in person hours and effort. Miscellaneous expenses were shown to be equivalent in quantity and cost. The initial effort of setting up and developing the MMIs and additional rooms required is offset by the improvement of their interview’s psychometric properties. The MMIs are more cost-effective than the traditional panel based interview. In terms of interview hours the MMI is more efficient than the traditional interview, even if only one interviewer is used. In terms of observer hours the MMI is equivalent to the variant most commonly used by US medical schools (2 interviewers for 1 hour) but less efficient with shorter interview times or those performed with only one interviewer. However the fewer the interviewer the less tenable the interview process will be given the psychometric standards of rigour. The main limitation of the study is that the monetary value associated with assessors is highly context-specific across institutions. This study only shows findings from the McMaster institute.

Brownell et al. (2007) examined the feasibility of MMIs, as well as user reactions (see section on ‘studies investigating user reactions to MMIs’ for details). They reported that, compared to traditional interviews, MMIs did not cost any more, were carried out in a shorter time period, with fewer interviewers and less time required per interviewer.

### *5.6 Summary of implementation of MMIs*

Several studies investigated issues related to the implementation of MMIs, including the effect of security violations, coaching, and station duration on MMI performance, as well as the relative cost-effectiveness of MMIs as a selection tool. Reiter et al. (2005) found that knowledge of the MMI stem questions two weeks in advance of the MMI did not improve performance scores, suggesting that security breaches would not be detrimental to MMI use as a selection tool. Griffin et al. (2008) reported that coaching did not improve MMI scores, and found that practice effects (when a candidate had experienced an MMI the previous year) resulted in small improvements in MMI scores on stations using the same or similar content as the previous year, but not on stations with novel content. These studies suggested that, although security breaches should not be detrimental to MMI use as a selection tool, the questions and content should be sufficiently varied across years to avoid practice effects. Dodson et al. (2009) examined the effect of MMI station duration on scores, and compared the scores of the same interviewer at 5 minutes and 8 minutes. Reducing duration from 8 to 5 minutes had a negligible effect on reliability and did not significantly affect applicant ranking, but could result in resource-savings (e.g. interviewer time). However, significant mean differences in scores were found, with lower scores at 5 minutes compared to after 8 minutes. Furthermore, using scores made by the same interviewer would likely minimise differences caused by MMI duration. When the cost-effectiveness of MMIs was compared to traditional panel interviews, Rosenfeld et al. (2008) concluded that, overall, MMIs are more cost-effective. Initially, MMIs demand greater effort to develop and prepare, as well as more rooms, but they are more efficient with respect to person hours and cost, particularly given the typical psychometric benefits of MMIs over traditional interviews. Brownell et al. (2007) also supported the administrative feasibility of MMIs: they cost no more than a traditional interview, did not require more interviewers, and were completed in a shorter period of time.

### *5.7 Studies investigating user reactions to MMIs*

Humphrey et al. (2008) carried out a study to assess candidates’ and interviewers’ perceptions of the use of a MMI for selection of senior house officers to a UK regional paediatric training programme. Candidates to a UK regional paediatric training programme n = 72 (86% of which were International Medical Graduates (IMGs), 75% were Foundation Programme Year 2(F2s), 10% were Senior House Officers, 4% were F1s and 6% trust grades doctors and interviewers (n = 15) who were experienced consultants. The questionnaire study looked at candidates and interviewers who completed an anonymous questionnaire directly after completing or observing an MMI. Questions asked about the fairness of the MMI and responses were on a 1-6 point Likert scale. Data was analysed using Mann Whitney and Kruskall-Wallis tests for comparisons. Both candidates and interviewers felt that the MMI was a fair process. IMGs preferred the format more than the UK graduates. Interviewers reported that MMIs were a better format than traditional interviews (mean score 4.8) and were a more reliable process (mean 4.4). Candidates and interviewers agreed that the MMI format was reliable and fair way of selecting candidates.

The main limitation of the study was that it was based on opinions of the participants and not fact. The study highlights that it is important that the all stakeholders are confident in the process of selection and believe it to be a reliable tool.

A study carried out at the University of Calgary in Canada by Brownell et al. (2007) examined the acceptability of the MMI by applicant and interviewers, and how well the applicant and interviewers felt they were prepared for the process. Interviewers were introduced to the MMI in a 2 hour didactic session several weeks before the interview days and given their assigned scenarios 48 hours before the interview day. Applicants were made aware of the process via the Admissions Office Web Page and written information. The population of the study was 281 applicants and 81 interviewers who participated in the 10 station MMI over a two day period. Ninety-eight and half percent of applicants and 91% of interviewers completed evaluations. An evaluation questionnaire (using a 5 point Likert scale) was distributed at the end of each of two MMI sessions to applicants and interviewers. The findings showed that interviewers felt well prepared for the MMI by the orientation session held the day before the MMI (mean 4.38) and by the information received about the station (mean 4.59). They reported having adequate time to assess the applicants (mean 4.08), considered the MMI a fair assessment (mean 4.28) and that the scoring sheet allowed them to differentiate among applicants (mean 3.97).98.6% were willing to participate in MMIs in future years.91.2% of applicants reported that they were participating in MMIs for first time. The information received beforehand prepared them (mean 4.06), they considered the MMI to be free of gender and cultural bias (means of 4.67 and 4.58), and had sufficient time to present ideas at the stations (mean 3.64). The MMI approach was found to be very acceptable to applicants and interviewers, for example with regard to perceived fairness, time to present/assess ideas, and good preparation beforehand. Interviews were carried out in a shorter time period than traditional interviews, with fewer interviewers and less time required per interviewer. However the limitations of the study were that only 18% of applicants compared to 76% of interviewers wrote comments on the evaluation form, which may have reflected applicant fatigue. In addition the study was only conducted at one site. The study shows the feasibility of conducting MMIs, with interviews conducted over a shorter time period than traditional interviews and less time required per interviewer. It shows their acceptability to participants and interviewers, with positive perceptions of prior information, fairness and time.

Eva et al. (2004a, 2004c) reported the results of post-MMI surveys that provided positive feedback on the process from both interviewers and candidates (see section on ‘studies investigating reliability of MMIs’ for details). In both studies, candidates found MMIs to be no more anxiety-provoking than a traditional interview.

### *5.8 Summary of user reactions to MMIs*

User reactions to MMIs have generally been positive. Both interviewers and candidates perceived MMIs to be fair (Humphrey et al., 2008; Brownell et al., 2007). Candidates have reported they regard MMIs to be free from gender and cultural bias (Brownell et al., 2007) and no more anxiety-provoking than a traditional interview (Eva et al., 2004c). Interviewers also felt that the process was reliable and they preferred the MMI format compared to traditional interviews (Humphrey et al., 2008), although the need for interviewer training for MMIs was highlighted in one study (Eva et al., 2004a). Brownell et al. (2007) also reported that interviewers felt they had sufficient time to assess candidates, and candidates felt they had sufficient time to present their ideas, although some candidates surveyed by Eva et al. (2004a) would have preferred longer (10 minute) stations.

### *5.9 Discussion and conclusion*

Multiple mini interviews are a relatively new method of selection and were developed in Canada from where the majority of evidence comes.

It is important that all users of a selection method have confidence in it; both interviewees and interviewers felt confident in this method, reporting that MMIs reduce cultural and gender bias. In addition interviewees felt they had sufficient time to present their ideas and interviewers had sufficient time to assess candidates. Interviewers also reported that the process was more reliable and preferable over traditional interviews.

Typically MMI stations each lasted 8 minutes with one rater observing the candidate. However the number of stations varied between 5 and 14. The studies found that increasing the number of stations to 14 had a greater impact on reliability. MMIs have several strengths: they can be designed to best fit relevant competencies, e.g. teamwork, communication skills. Studies suggest that MMIs are a reliable measure of non-cognitive attributes such as empathy, cultural sensitivity, advocacy and communication skills which are all important attributes for a doctor which can be observed in a specific context. Inferences may be made that MMIs help to protect patient safety as candidates are being observed and rated on doing a work sample.

Eva et al. (2004b) and Reiter et al. (2007) found MMIs to be a better predictor of performance on OSCEs than more established methods of selection (e.g. GPAs and traditional interviews). However more research needs to be carried out as to whether MMIs support predictive validity. MMIs cannot be coached for as there is not one formulaic answer (as in one of the weaknesses of white space) and therefore security breaches in the selection process would not affect the performance of the candidate (Reiter et al., 2005 and Griffin et al., 2008).

MMIs are initially costly and time consuming to set up in terms of person hours and expense (e.g. development of the blueprint of competencies and scenarios, training the raters/interviewers and hiring the number of rooms required for the stations). However the value in terms of the reliability and validity over more traditional unstructured interviews may outweigh these initial costs.

## 6. National examinations

It is assumed that the quality of UK medical school education is high; however there is no national exit exam to confirm a common standard. The UK higher education system relies on an 'external examiner' system which involves academics from different institutions attending and reviewing the assessments made to ensure comparable standards. A study comparing OSCE scores set by five different medical schools found that the pass rate set at each station varied widely, such that a student may pass at one medical school but fail at another (Boursicot et al*.,* 2006). This study raises a concern that the range of standards may in some cases be low. A finding which is a cause for concern and may indicate a need for a national exit exam.

Most of the literature on national exit examinations comes from the United States and from Canada. A review by Balentine, Gaeta and Spevack (1999) found that performance in medical school did not correlate with achievements in medical residency posts, but that academic performance in medical school and performance in board certification exams does have a positive correlation. In a review paper Berner et al. (1993) report that many studies have examined the relationships between students' performances on the National Board of Medical Examiners (NBME) Part I and Part II examinations and their postgraduate clinical performances. Berner et al*.* reported most studies have yielded a low to moderate correlation of the NBME I, II and III scores with medical school and residency performances. They advise against using the new United States Medical Licensing Examination (USMLE)[1] as the sole criterion for final resident section as clinical performance is multidimensional and the USMLE is not a sufficient measure of clinical performance. Lee et al. (2008) comments that USMLE web site states "the primary focus of USMLE is for licensure decision", however, despite this the USMLE scores have been used for selection.

A recent BEME review (Hamdy et al., 2006) aimed to assess the value of measurements obtained in medical schools in predicting future performance of medical practice. They were able to combine data from 19 of the 38 studies identified to conduct a meta-analysis. They reported that the highest correlation between predictor and outcome was between the National Board Medical Examination (NMBE) Part II and NMBE Part III (r = 0.72) and the lowest between NBME I and supervisor rating during residency (r = 0.22). The review draws our attention to the complex nature of measuring performance in practice, of competence 'what a doctor is capable of doing' under test conditions and performance 'what he or she does do in day to day practice'. The diagram below taken from Hamdy et al. but based on

[1] The USMLE Part 1 examination is a standardized national examination that returns for each examinee a three digit score. Three digit scores are equated across time and exam form, such that identical three digit scores – regardless of the year in which the examination was taken – imply equivalent levels of performance.

Suggestion that USMLE test results could be biased by medical school characteristics (not proven) and curricula designed to teach specifically to the format and content of the USMLE exam. Finally the USMLE exams are not designed to predict future performance of physicians. (Dirschl et al., 2006)

Miller (1990) highlights the relationship of 'learning outcomes' to 'practice outcomes'. Students can be assessed at any of the three learning outcomes measures but most of the student assessments are based on the lower two levels 'knows' and 'knows how'.

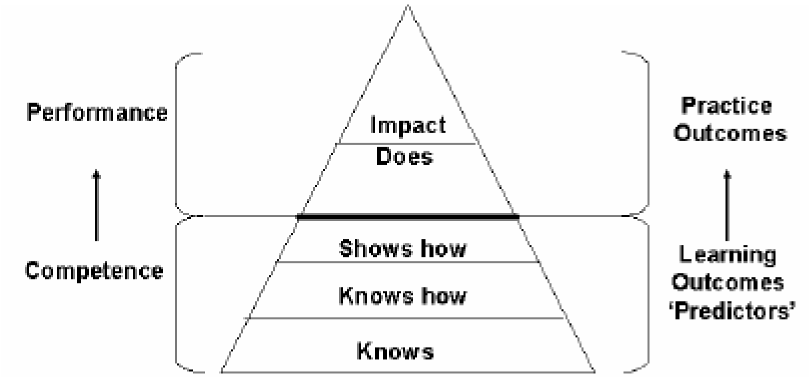

Fig 1: Conceptual relation between assessment in medical education of learning outcomes 'predictors' and "practice outcomes".

Towards the end of medical school the third layer becomes more important as graduates prepare for practice. Recent research by Illing et al. (2008) found that this third layer was where the gap occurred from leaving medical school to starting work as a foundation doctor.

### *6.1 Studies that report national examinations scores as a predictor with later examination scores*

The meta-analysis by Hamdy et al. (2006) highlights that scores on exit exams are at least moderate predictors of scores on later exams. The highest reported correlation in the meta-analysis was between NBME II (predictor) and NBME III (outcome) for two studies (Market, 1993; Hoyat et al., 1993) (r = 0.72). The same two studies also showed a moderately good correlation between NBME I and NBME III (r = 0.59). A moderately good correlation was reported for a meta-analysis of five studies (Case, 1993a,b,c; Boyse, 2002; Sosenko, 1993) that correlated NMBE I scores as predictor with American Board of Speciality Examination (r = 0.58) and again a moderately good correlation was reported for a meta-analysis of three of these five studies (Case 1993a,b,c)that correlated NMBE II scores as predictor with American Board of Speciality Examination (r = 0.61).

The aim of a study by Black at al. (2006) was to identify any relationship between scores on the US Medical Licensing Exam (USMLE) and the Orthopaedic In-Training Exam (OITE) over time. USMLE is used in selection and covers similar areas to OITE. The study involved 23 Penn State department of orthopaedics and rehabilitation residents across eight years. There were small numbers in each cohort – e.g. only 12 available for analysis in year 4. The study involved looking for correlations between USMLE step 1 and step 2 (standardised three-digit scores), and OITE centiles for each year in training between 1993 and 2000. A moderate correlation was found between USMLE step 2 scores and performance on the OITE, increasing in value through years. The sample size was small and the authors suggest this may account for the lack of correlation with USMLE step 1.

The aim of a study by Thordarson et al. (2007) was to evaluate the agreement among evaluators regarding the importance of various selection criteria and to assess whether any objective measures of performance during or after residency correlate with the initial rank list or the residents' USMLE Part 1 scores. The study involved 46 residents (3 classes of 12 and 1 class of 10) from an orthopaedic surgical department, Los Angeles, USA.

Four members of the orthopaedic department interviewed four classes of residents. Interviewers were asked to rank the residents against selection criteria in order of importance for their personal decision-making process. The selection criteria incorporated nine elements: interview, letters of recommendation, rotation evaluations of performance, and any calls made or received on behalf of the applicant, personal statement, United States Medical Licensing Examination(USMLE) Part 1 Scores, medical school grades and/or Alpha Omega Alpha (AOA) honour status (combined as some schools do not have an AOA chapter), research experience as a medical student, medical school attended, other activities such as employment or recreation, sports and volunteer work and gender and/or ethnicity.

Residents' performances were assessed numerically with their in-service scores (fourth year Orthopaedic In-Training Examination (OITE) scores) and their American Board of Orthopaedic Surgery (ABOS) Part 1 scores. Performance was ranked 1-12 (or 1-10 for the class of 10) for each of the four years. Additionally residents were ranked within a "Best Doctor" award voted by all residents (average 63% response to ballots). The strongest correlation was between USMLE score ranking and the ABOS scores ranking (R2 = 0.59). Fair or poor correlations were found between the residents' initial rankings, rankings on graduation, and their USMLE, ABOS and OITE scores. Interviewers did not agree on their rankings of residents at graduation. The USMLE scores showed a correlation with the interviewer selection, indicating that knowledge of the USMLE biased the interviewers.

The aim of a study by Stacy et al. (2005) was to identify what predicts academic and clinical success for international dental students. The study involved 183 international dental students at Loma Linda University, USA. The study examined application variables (including demographics, National Board Part I (science) and II (clinically focused) academic results, and Test of English as a Foreign Language, dexterity tests and interview scores) with final academic grade and final clinical score on a postgraduate programme designed to ensure international graduates meet required standards.

Dexterity tests and interview are conducted following initial selection from paper application. Part I is normally taken after two years in dental school, and Part II just before graduation, so both are prerequisite for international applicants. Multiple regression analyses identified the National Board Part 1 and Part II as significant predictors of academic grade (r = 0.39.0.442). Academic and clinical grades were found to be highly correlated (r = 0.7). Regression found Part II and dexterity to be the best predictors of Academic grade and Clinical Grade.

Bell et al. (2002) carried out a study which aimed to determine whether the United States Medical Licensing Examination (USMLE) scores correlate with in-training examination results from all four years of residency training and to compare the selection criteria (USMLE scores, medical school grades, and faculty interviews of applicants) to an overall assessment of residency performance. This comparative study looked at scores from the USMLE step I and II Obstetric and Gynaecological residents (who had graduated or trained at a particular Hospital in America between 1995 and 1999 (n = 20)) and compared their scores on in-training examinations. Faculty then rated their cognitive and non-cognitive clinical performance. These scores were then compared with criteria of their medical school performance. The analysis the authors used was linear regression. Results from the study showed that USMLE scores positively correlated with in-training examination scores. However USMLE scores, honour grades in student clinical rotations and student interview scores did not correlate with the faculty evaluation of resident performance. In conclusion selection criteria that are based on other medical school achievements do not necessarily correlate with overall performance as residents in Obstetrics and Gynaecology. The main limitation of this study is that the results cannot be generalisable as the data is based on one site, in a single specialty and a small sample size.

### *6.2 Studies that report national examination scores predict later examination scores but have low predictive validity for clinical performance*

Hamdy et al. 's (2006) systematic review discussed above found moderately good correlations for the predictive validity of national exams, but low predictive validity for clinical performance using supervisor ratings as the outcome (see above for details).

The aim of a study by Brothers and Wetherholt (2007) was to examine the relationships between selection criteria for surgical residents and subsequent performance, in order to assess the effectiveness of the selection system. The study involved 152 applicants interviewed for the programme and 26 residents accepted into the program. It was conducted at the university-based surgery residency programme at the Medical University of South Carolina, USA. The method involved a series of correlations to assess which assessments were more closely related to core competency based clinical rotation performance ratings. The results showed that medical school GPA and USMLE scores correlated negatively with clinical performance, but positively with two in-training exams (ABSITE and ABS QE). Residents who required remediation during training actually carried higher USMLE and GPA grades ($p<0.05$). The study reported that non-cognitive factors; personal characteristics and reference letters correlated most strongly with subsequent performance, however the limitations of the study were that surgical faculty provided the ratings both at interview and after each rotation.

### *6.3 Studies that report national exams have a low predictive validity for subsequent exams*

The aim of a study by Gunderman et al. (2000) was to determine whether the National Board of Medical Examiners (NBME) examination scores have any value in predicting the subsequent examination performance of residents. The study took place in the department of radiology, Indiana, USA. This was a retrospective study of NBME scores and the American Board of Radiology (ABR) scores for 99 radiology residents over a period of 10 years. The NBME did not show any statistically significant value of NBME examination scores in predicting eventual performance on the ABR written and oral examinations. The future success of students on radiology exams could not be predicted from the ranking of students on the basis of their NBME scores.

The sample size may be too small to show a relationship. The numbers who failed to pass the ABR were very low, restricting the statistical analysis. It is possible that the department selected the higher ranking students and those who were lower ranking students were missing from this analysis, biasing the results. Lower ranking students on the NBME may have shown lower ability to pass the ABR.

### 6.4 Studies that report national exams as low on predictive validity for other exams and clinical performance

The aim of a study by Turner et al. (2006) was to compare the predictive ability of a composite selection tool, the quantitative composite scoring tool (QCST). The QCST consists of: medical school reputation, class rank/AOA status, basic science honours grades, junior clinical clerkship honours grades, Mayo orthopaedic clerkship grade, USMLE Part 1 percentile score, undergraduate grades, graduate school degrees, letters of recommendation, miscellaneous/ extracurricular activities. The QCST was compared to three other predictor variables: USMLE-1 scores, AOA and the junior year clinical clerkship (JYCC) grades. The study involved 64 residents who matriculated into the Mayo Orthopaedic Residency Program, Rochester, USA. The study was testing the predictive validity of the USMLE part 1, AOA status, JYCC and the QCST against four residency outcome assessments. The outcomes included three standardized assessments: the orthopaedic in-training examination scores (OITE), the American Board of Orthopaedic Surgery (ABOS) written and oral examinations and an internal outcomes assessment of satisfactory chief resident associate (CRA) status. Collectively the QCST score was the strongest predictor for all of the three standardised outcomes (p,0.01). The authors conclude that the quantitative composite scoring tool can be more effective in predicting residency outcomes as compared to individual predictors such as USMLE 1 scores, AOA status or junior years clinical performance.

### 6.5 Studies that report scores on national exams were low predictors of later clinical performance

The Hamdy et al. (2006) meta-analysis on seven studies (Market,1993;Smith,1993; Kahn et al., 2001;Yindra,1998a,b; Alexander et al., 2000; Paolo et al., 2003); found a low correlation between NBME II and supervisor rating (r = 0.27). The same seven studies plus two more (Borowitz, 2000; Paolo et al., 2003) were then entered into another meta-analysis and found an even lower correlation for NBME I and supervisor rating (r = 0.22).

The aim of a study by Park et al. (2006) was to evaluate the ability of pre-dental and preclinical benchmarks to predict clinical performance in a cohort of 84 dental students at Harvard School of Dental Medicine, USA. The predictors were overall GPA, science GPA, dental admissions test (DAT) and performance on National Board dental examination. The outcome variables were measures of student performance and productivity during the clinical years (the four disciplines were: operative, major restorative and removable and fixed prosthodontics). The results of the study revealed a lack of any statistical associations between the predictors and the clinical outcomes. The authors suggest the sample size may have been too small to show an effect or that the skills set required to show clinical outcomes may vary significantly from the skill set required to perform well on standardised examinations.

The aim of a study by Wood et al. (1990) was to compare objective measures of the radiology resident applicant performance such as NBME scores with non-objective measures such as conscientiousness and interpersonal skills in their prediction of resident performance. The study involved 30 (49% participation rate) applicants who applied for a radiology residency at the University of IOWA in 1983. Applicants who applied for the radiology residency were examined. As part of the application process the applicants supplied their NBME scores (part 1 and 2) class rank, AOA membership status, the number of research projects that they had been involved in during medical school, and their number of publications.

Applicants participated in traditional unstructured selection interviews that were given a mean score on the subjective evaluation of four faculty interviewers. Applicants were also interviewed by at least one current resident who used the same interview format. None were blind to the applicants' AOA status, NBME scores or class rank. The Accomplishment Interview (AI) was administered at the time of interview. Prior to their visit applicants were sent a questionnaire that asked them to describe past situations in which they had best demonstrated the qualities: interpersonal skills, conscientiousness, recognition of limits, confidence and curiosity. During the visit applicants met with an interviewer to discuss and clarify their questionnaire responses. Interviews were scored independently using a behavioural benchmark.

NBME scores failed to predict residents' performance and did not positively correlate with any of the criterion variables. NBME scores correlated negatively with several criterion variables: manual dexterity as measured by the radiology evaluation, motivation as measured by the radiology evaluation, interpersonal skills as measured by the Behavioural Observation Scale (BOS), and confidence as measured by the BOS. NBME scores were positively correlated with membership in Alpha Omega Alpha (AOA) (r = .67), but AOA did not correlate with any performance measures.

The BOS total score and the standard radiology evaluation had a positive correlation of 0.42 (P<0.05). The interpretive skills subscale of the standard evaluation was correlated with conscientiousness (r =0.57;P<0.001) the confidence (r=.34; P<.04) and the interpersonal (r=.59; P< .001) subscales of the BOS. The BOS curiosity scale was moderately correlated with the motivation subscale of the standard evaluation. (r = .36; P<.05). There was a moderate correlation between BOS interpersonal subscale and the manual dexterity subscale of the standard evaluation (r =.37; P< .05).

Wood et al. argue that NBME scores were designed to measure performance on the content of medical school curricula and not intended as a assessment of preparation for residency. The aim of a study by Papp et al. (1997) was to examine whether admission variables (based mainly on exam scores and GPAs) to a general surgery residency programme could predict general success during residency. The study involved 17 faculty members from the surgery department who then rated 39 general surgery residents. The study was conducted at the University of Louisville School of Medicine, USA. The results showed that there was no relationship between the exam based admissions data and final evaluation. There was no significant correlation between rank order assigned before residency and rank order based on performance following residency (r=.27). No significant differences were found between residents ranked in the top and bottom ten. ABSITE score in the first year of residency could not be predicted from medical school class rank. This study highlights that ranking based on exams does not predict later ranking based on faculty ratings of residents' performance.

The aim of a study by George et al. (1989) was to study the effectiveness of an intern selection committee's decisions in comparison to future performance. The study involved 161 interns (20 interns per year over a 13 year period. The interns entered a 3 year postgraduate training programme in internal medicine, Columbus, USA. The selection criteria studied included class rank, grades in four major clinical clerkships, scores on the NBME Part 1, AOA membership and ranking by the intern selection committee. Results showed that only the ranking by the intern selection committee was correlated significantly with subsequent house-staff performance. However two items from the intern selection folder were highly correlated with the intern selection committees ranking. These were NBME I scores (p<.01 Spearman Rank) and membership of AOA (P =0.01 Kriuskal Wallis), highlighting that the committee were influenced by these factors more than grades in the four major clinical clerkships.

### *6.6 Studies that report on the Objective Structured Clinical Examination (OSCE)*

The Hamdy et al. (2006) systematic review also conducted a meta-analysis on five studies (Smith, 1993; Kahn et al.2001; Vu et al., 1992; Rutala,1992; Wilkinson, 2004) to determine whether OCSE scores predicted supervisor ratings. The correlation was low (r = 0.37). Hamdy et al. suggested this weak correlation, although statistically significant, may be the result of the supervisors not measuring the same constructs as assessed by the OSCE in the undergraduate department.

The aim of a study by Rifkin et al. (2005) was to determine whether performance with standard patients as measured by OSCE correlated with USMLE step 1 and 2. The study involved 34 postgraduate year one internal medicine house-staff in a community teaching hospital in the USA. The study method involved testing for a correlation of USMLE Step 1 and Step 2 with standardised patient scores from an OSCE (history taking, physical examination, and interpersonal skills). The two USMLE steps correlate well (.65), but OSCE scores correlate low with both (R = 0.20 for step 1, .09 for step 2). The findings indicate that USMLE does not predict clinical performance, or the OSCE was an inadequate measurement of clinical performance.

### *6.7 One study that reports scores on national exams were predictors of later clinical performance*

The aim of a study by Tamblyn et al. (2007) was to assess whether patient-physician communication examination scores in the clinical skills examination predicted future complaints in medical practice. The study involved a cohort of 3424 doctors who took the Medical Council of Canada clinical skills examination between 1993 and 1996 and were licensed to practice in Ontario and Quebec. Participants were followed up until 2005 (the first 2 to 12 years of practice). Multivariate Poisson regression was used to estimate the relationship between compliant rate and scores on the clinical skills exam and traditional written examination (taken after medical school).

The results showed that 17% of the doctors had at least one retained complaint and 81.9% were for communication or quality-of-care related issues. The patient-doctor communication score in the clinical skills exam remained significantly predictive of retained complaints (p, .001 with scores in the bottom quartile explaining an additional 9.2% (95% CI, 4.7%-13.1%) of complaints. Communication and clinical-decision making ability were important predictors of future complaints to regulatory authorities. The authors suggest that current exams could be modified to test these attributes more efficiently.

### *6.8 Discussion of strengths and weaknesses of using national exams as part of selection*

The United States licensing exams have been used widely for the selection of residents. The data presented here has highlighted in particular the evidence from the BEME systematic review by Hamdy et al. (2006). that similar types of exams such as the NMBE I, II and III are moderately predictive of later exam scores including some speciality examinations. Three more recent studies by Black et al. (2006), Thordarson et al*.* (2007) and Stacy et al. (2004)

all add further support. We identified one study that contradicted this general finding, Gunderman et al. (2000); however the numbers failing the later ABR exam were small, restricting the analysis. Turner et al. 's (2006) study showed that a composite score that included the USMLE, AOA and junior year clerkship grades was a better predictor of later exam scores than USMLE alone.

Hamdy et al. (2006) reported on nine studies that had a low combined correlation between exams and clinical practice as measured by the supervisor rating. Four additional studies reviewed here, (Park et al., 2006; Wood et al., 1990; Papp et al.1997; George et al.1989) all reported that exams were not related to later clinical performance.

Hamdy et al. (2006) also reported on five studies to examine the predictive validity of OSCEs to determine clinical performance. Although a correlation was reported, the correlation was weak, and a similar finding was reached by an additional study we reviewed (Rifkin et al.2005). The Tamblyn et al. (2007) study was the only study that found a link between exam scores and later practice. The study reported on one aspect of the Medical Council of Canada clinical skills examination with particular reference to the patient-physician communication examination scores which were found to be predictive of future complaints in medical practice.

### *6.9 Conclusion*

There is evidence that the strength of national examination scores is that they are moderately predictive of later examination scores. The main weakness of exams is that they have low predictive validity for clinical practice. OSCEs are more practice orientated but as yet they too show low predictive validity. If selection intends to consider future suitability for clinical practice then exams and OSCEs may not be the most relevant assessments available.

Returning to the hierarchical pyramid discussed earlier, national exams have greater predictive validity for knowledge located at the base of the pyramid but their weakness is the low predictive validity in relation to practice.

## 7. Academic grades

Most Western medical schools look for some degree of academic excellence as predicted by school grades and pre-admission examinations (Glick, 2000). In the UK selection for medical school is traditionally based on predicted or actual school-end A-level results.

This section of the literature review focuses on the relationship between academic performance before admission to the undergraduate or postgraduate medical education programme and subsequent performance.

### *7.1 Grade point average (GPA): positive findings regarding future academic performance*

Measures of academic performance used have included overall undergraduate grade point average (GPA), undergraduate science GPA and school-leaving grades. A previous review has reported that 'The evidence is overwhelmingly clear that pre-admission academic grades predict subsequent in-course academic performance in all professional disciplines' (Salvatori, 2001, p.161). Another review (Ferguson et al., 2002) reported that previous academic performance (including the medical college admission test, A levels and grade point average) does predict achievement in medical training, accounting for 23%of the variance in performance in undergraduate medical training but only 6% of that in postgraduate medical training.

Iramaneerat (2006) conducted a study involving 223 medical students entering the Faculty of Medicine Siriraj Hospital (Thailand) in 1997 to investigate whether high school grades can predict medical school grades after controlling for the effects of demographics and entrance exam scores. Hierarchical multiple regression analyses were used to predict medical school grades using demographics (age and gender), entrance examination scores and high school GPAs as predictors. Dependent variables were grades from three study levels of the Faculty of Medicine curriculum: premedical, preclinical and clinical. High school grades provided significant prediction only for premedical grades (i.e. short-term academic achievement), and only for students who studied in a traditional (rather than expedited) high school curriculum. The type of entrance examination taken and the type of high school curriculum studied were significant predictors of medical school grades at every level. Long-term academic achievement could be better predicted by academic orientation, commitment to medical study and demographic traits than measures of cognitive abilities. Limitations of the study included limited population generalisability, as students with low entrance examination scores are not admitted into the medical school; limited predictive power due to the maturation effect (clinical grades were not available until six years after medical school entrance); and a limitation in the outcome variables, focusing only on an academically dependent outcome and not on non-academic skills.

Several studies reviewed here have positive findings regarding GPA and future academic performance. One reason why GPA has shown predictive utility may be because it is an average of several measures of applicants' ability, based on an extended period of time. Since it is an average, variance due to extraneous factors is more likely to have been cancelled out (Kulatunga-Moruzi & Norman, 2000b).

The following study reports on *perceptions* of the predictive ability of academic performance. Swide et al. (2009) aimed to study the perceptions of anaesthesiology resident programme directors about the value of the Medical Student Performance Evaluation (MSPE) in predicting successful residents, and to identify which sections were and were not predictive. The MSPE has replaced the traditional Dean's letter as a component of application to a residency program. An online 10-item survey, with both quantitative (4-point Likert scale) and qualitative sections, was sent to 115 United States medical-school based anaesthesiology residency programme directors. The response rate was 38%. Sections of the typical MSPE that were perceived as predictive (rates of 60% or higher) were a) academic history summary, b) academic progress, c) academic ranking and d) comparative clinical performance. Sections perceived as not predictive (ratings below 60%) were a) unique characteristics, b) pre-clinical comparative performance, c) professional behaviours/attitudes compared to classmates, d) summary statement and e) Appendix E (information specific to the medical school). The strongest theme emerging from the qualitative data was a desire for the MSPE to indicate candidates' rank. The authors suggest that there is a trend by programme directors to rely on the most objective sections of the MSPE in the selection of anaesthesiology residents, such as class rank, academic performance relative to others, need for remediation, and overall performance relative to peers. The study suggests that there is a lack of confidence among anaesthesiology programme directors in the value of the MSPE assessment of professional behaviours, despite evidence that professionalism is a key indicator of success in residency programs and later practice.

Hall et al. (1992), discussed below, found no correlation between GPA scores and dean's letter of recommendation which represents an opinion based on academic performance as to how likely a student is to perform well in their future residency.

Evans and Wen (2007) conducted a cohort study to investigate the extent to which Medical College Admission Test (MCAT) sub scores predict the overall academic performance of osteopathic medical students. The study subjects were 434 osteopathic medical students of the Oklahoma State University College of Osteopathic Medicine in Tulsa, USA. The study found that total undergraduate GPA was the most significant predictor (ß = .13-.33) in overall student academic performance as measured by basic GPA, clinical GPA, total GPA, and Comprehensive Osteopathic Medical Licensing Examination-USA (COMLEX-USA) Level I and Level 2 scores. MCAT sub-scores were less predictive of overall academic performance (ß = -.01-.21). A limitation of the study stated by the authors was that it was unclear whether the results were unique to osteopathic medical education.

Baker et al. (1993) carried out a retrospective case review of CARCS (computer-assisted resident selection) data that revealed some of the strengths and weaknesses of CARCS and of the process of resident selection in general. CARCS is a database application developed in 1983 at the Department of Anesthesiology of the Medical School of South Carolina. CARCS database files were analysed for two recent years (1990-91 and 1991-92) and two consecutive years five years earlier (1985-86 and 1986-87). Simple averages and percentages were derived for two groups: 1) the entire pool of residency applicants being considered and 2) the residency candidates who actually matched with the program. Undergraduate GPA (3.53) tended to run about half a point higher than the medical school GPA (3.12). There were consistently higher average dean's letter and other letters of reference points for matched residents (3.85) than for corresponding pools (3.42). Interview scores averaged for matched residents (4.11) were also higher than for the corresponding candidate pool (3.77). The authors conclude that the parameters of undergraduate and medical school GPA, MCAT, NBME scores and class rank are not definitively associated with successful match, likely because better-than-average grades are almost a given. Quality of reference appears to be the most important component of pre-interview score in determining match success, and the personal interview is the other important factor. Limitations of the study were variation in data availability, for example medical school GPA was unavailable for over half the applicants, recent modification to MCAT scoring, and the NBME being phased out to be replaced by the USMLE.

Eva et al. (2004b), discussed earlier, found that GPA was the best predictor on multiple choice exams measuring medical knowledge (Personal Progress Inventory (PPI)) (ß = 0.54, p<.05), followed by an autobiographical submission (ABS) (ß = 0.45, p<0.5). Whereas the ABS scores had a relatively strong relationship with PPI performance early in the programme but correlations rapidly declined, the GPA did not predict performance on the first PPI, but steadily improved in its ability to predict later PPI performances. The finding that GPA was the best predictor on medical knowledge tests replicates the findings of Kulatunga-Moruzi and Norman (2002a) regarding GPA as a predictor of performance on cognitive outcome measures. They carried out a study to examine the utility of several cognitive and non-cognitive criteria used in the admissions process in predicting both cognitive and non-cognitive dimensions of the licensing examination of the Medical Council of Canada (LMCC). The study subjects were 97 students who enrolled in the McMaster University (Ontario, Canada) undergraduate medical programme in 1993 (analysis based on between 52 and 97 students due to some missing data). Performance on the LMCC Part I (assessing knowledge in six areas of medicine and clinical reasoning) was significantly correlated with overall undergraduate GPA (r = 0.327, p<0.001), undergraduate science GPA (r = 0.446, p<0.00001), MCAT verbal sub-score (r = 0.318, p<0.05) and MCAT total score (r = 0.329, p<0.05). Performance on LMCC Part II total and on the problem-exploration and data acquisition component were significantly correlated with only overall undergraduate GPA (r = 0.251, p<0.01, r = 0.274, p<0.05, r = 0.21, p<.05 respectively). The communication skills component of the LMCC Part II component was significantly correlated with the MCAT verbal sub-score *(*r = 0.429, p<0.005) and the personal interview score (r = 0.239, p<0.05). Limitations of this study are that it is a single site study, with data on one year's admissions only. Reiter et al. (2007), discussed earlier, similarly reported that only the uGPA predicted progress scores (PPI) and was statistically predictive of performance on the multiple choice medical knowledge exams (standardised ß = 0.38, p<0.05), whereas MMIs predicted OSCE performance, clerkship performance and the professional/non-clinical components of a licensing exam.

Utzman et al. (2007), in a study involving 3365 students from a nationally representative sample of twenty physical therapist programs in the USA, which aimed to assess whether admissions data could be used to estimate the risk of failing the National Physical Therapy Examination (NPTE) and whether the degree of prediction varied by program, found that uGPA, quantitative and verbal Graduate Record Examination (qGRE and vGRE) scores, and race or ethnicity were independent predictors of NPTE failure. Controlling for other variables, the odds of failing the NPTE (a 200 multiple-choice question, standardized examination administered by computer) increased 12% for each 0.10 decrease in GPA, 6.6% for each vGRE drop of ten points and 3.5% for each qGRE drop of ten points. Ethnicity also predicted NPTE failure. However, the study only predicts failure on a test rather than high levels of performance. The between-programme logistics regression model (which included programme as a predictor) accounted for 28% of the variance in the odds of NPTE failure and there remains a considerable amount of unexplained variance (72%).

### *7.2 Positive findings on longer-term effects of GPA and entry level*

McManus et al. (2003) conducted a prospective cohort study with follow up after 20 years by postal questionnaire to assess whether A level grades (achievement) and intelligence (ability) predict doctors' careers. The four outcome measures were dropout, career progression, research output, and stress, burnout and satisfaction with medicine as a career. The study participants were 511 doctors who had entered Westminster Medical School (London, UK) as clinical students between 1975 and 1982 and were followed up in January 2002. They had taken a standard intelligence test (AH5) on entry to Medical School, measuring verbal and reasoning ability (Part I "verbal") and spatial ability (Part II "spatial"). The follow-up questionnaire asked about career, qualifications, interests and personality. Stress was assessed with the General Health Questionnaire (GHQ-12) and an abbreviated version of the Maslach Burnout Inventory (aMBI) with additional questions on satisfaction with medicine. Forty-seven doctors were no longer on the Medical Register. They had lower A level grades than those who were still on the register (p<0.001) but not lower AH5 scores. Of the 464 doctors on the register, there was a 73% response rate to the questionnaire. Non-respondents had lower AH5 scores but did not have different A level results. Hospital doctors had higher A level grades and AH5 scores than General Practitioners (GPs), each effect being significant after accounting for the other. A levels had significant effect on time taken to gain membership qualifications, choosing to become a GP or leaving the register (r = 0.376, SE 0.098, p<0.01). They did not predict diploma or higher academic qualifications, research publications, or stress or burnout. Intelligence did not independently predict dropout, career outcome or other measures. The authors conclude that A levels have validity in selection, with a validity coefficient of about 0.3, but state that results may not be generalisable to other examinations in other countries.

A study conducted in the Netherlands (Cohen-Schotanus et al., 2006) aimed to ascertain whether the GPA of school-leaving examinations is related to study success, career development and scientific performance. Admission to all medical schools in the Netherlands is determined by the use of a national lottery procedure weighted in favour of students with higher GPAs. The study involved 398 students admitted to the Faculty of Medicine at the University of Groningen in 1982 and 1983. The independent variable was the GPA from the school examination and the dependent variables were study success, career development and scientific performance (defined as publication of research papers and a PhD dissertation). Moderator variables were gender and cohort. The study found that a higher GPA significantly enhances study progress (coefficient -0.53), with significantly less time to graduation. Higher GPAs were associated with graduates qualifying in their preferred specialty and with higher scientific output. GPA scores had no effect on drop-out rate. Gender differences were found for study duration and scientific output, with women graduating earlier and publishing less. A strength of the study was its longitudinal data collection as graduates were interviewed about career development annually between 1993 and 2000. A possible limitation was that it was conducted in one medical school only.

Two studies reported on future impact of differences in ability at entry level. Paolo et al. (2006) conducted a retrospective study to compare the performance of alternate- and main-list students during and one year after medical school, assessing admission and performance measures for 1188 students matriculating from 1997-2003 at the University of Kansas School of Medicine, USA. Measures included MCAT scores, basic and clinical science GPAs, USMLE Step 1 and Step 2 scores, residency match information and residency director ratings. The study found that both the admission measures and performance of alternate-list students were generally lower than for main-list students, however the authors state that the differences were very small and probably not meaningful. The study does highlight that those starting medical school with slightly higher rankings continue to do very slightly better during medical school and during the first postgraduate year, and suggests that taking students with lower scores into medical school did not adversely affect overall performance during medical school and one year after graduation.

The study results may not be generalisable to medical schools that have substantially different admissions criteria.

McCarey et al. (2007) conducted a study to explore the prediction of academic performance by entry profile of a cohort of nurses. Study participants were 154 graduates from a three year nursing diploma at a Scottish University. A comparison was made between groups with different levels of entry qualification: no university-entry qualifications (i.e. access and standard grades), those with university qualifications (Highers) and those with tertiary qualifications. In Year one there was no statistically significant difference between groups in the two written assignments, although there was in examination results, with Group 3 having better scores. Group 3 also performed better at one piece of Year 2 coursework and at two pieces of year 3 coursework and one exam. There were slight effects of age, with older students tending to have higher marks than younger students, although this was statistically significant in one assignment only, and of gender, where differences were inconsistent. Whilst the study does give some indication that effects of entry level continue throughout the course, entry level is more variable for nursing than for medicine, so the results may have limited relevance.

Peskun et al. (2007) conducted a study to assess the effectiveness of medical school admissions criteria in predicting residency ranking four years later. The study sample consisted of five classes of University of Toronto, Canada, medical school students (1994-1998) who completed degrees and applied to positions in Family or Internal Medicine (n = 345 and 315 respectively) at the University in their graduating year (1999-2003). Predictive validity of application components was assessed by estimating the association between the components and the ranking of students by the two postgraduate residency programmes. The predictive validity of admissions criteria was assessed for both academic and non-cognitive medical school performance. Admission variables were academic measures (undergraduate GPA (uGPA), numerical Medical College Admissions Test (MCAT) score) and non-academic measures (personal essay, autobiographical sketch, letters of reference, interview). Medical school variables were OSCE score, Internal and Family Medicine clerkship grades and ward evaluations, and final medical school grade. Residency rank in Internal Medicine was correlated significantly with uGPA ($p = 0.0296$) and the admissions non-cognitive assessment ($p = 0.0394$), with a trend towards significance with MCAT. However, there was no relationship between GPA, MCAT and final application score and Family Medicine ranking, and a significant correlation with the admissions interview ($p = 0.0209$). Non-academic variables were correlated with rank in both residency programmes, but the differences between the correlations may reflect different emphases on academic and non-academic performance in the two programmes. Final grade in medical school was correlated significantly with GPA and non-cognitive assessment. A limitation of the study was that the subjects were a restricted cohort as they were all successfully admitted to the University of Toronto and thus had high academic and non-academic admissions scores.

### *7.3 Grade point average (GPA) as a predictor of performance*

A previous review states that ‘The relationship of pre-admission academic performance to clinical performance has been studied less often and is far less clear’ (Salvatori, 2001, p.162). A later review similarly states that ‘studies on the correlation of residency performance and grades in undergraduate and medical school have had mixed but not convincingly positive predictive results for residency performance’ (Lee et al., 2008, p.166).

The Hamdy et al. (2006) BEME systematic review conducted a meta-analysis of 11 studies (Market, 1993; Smith, 1993; Kahn et al., 2001; Yindra, 1998a & b; Alexandra et al., 2000;

Paola et al., 2003; Callahan et al., 2000; Pearson et al., 1998; Richards et al., 1962; Rabinowitz, 1989) of clerkship GPA and supervisor ratings. The combined correlation was low (r = 0.28).

Hamdy et al. (2006) also conducted a meta-analysis of five studies (Market, 1993; Loftus et al*.,* 1992; Yindra, 1998a & b; Richards et al., 1962) of pre-clinical GPA as a predictor of supervisor ratings. The combined correlation was low (r = 0.25).

Two studies reported positive findings regarding school/undergraduate grades and future OSCE performance. Lumb and Vail (2004) conducted a retrospective cohort study of 738 students who entered the University of Leeds School of Medicine between 1994 and 1997. They aimed to compare the relative importance of social, academic and application form factors at admission in predicting performance in a Year 3 objective structured clinical examination (OSCE). Analysis by multiple linear regression found that school-leaving (A level) grades and academic potential (Section C of the application form) were significant predictors of success in the OSCE (p = 0.002 and 0.05 respectively), other than for mature students who performed very well despite having poorer A-level grades (p = <0.001). The study also found that male and ethnic minority students showed a relative under-performance. Socioeconomic group and type of school attended were not found to affect performance. Overall, however, all the factors studied accounted for only 12% of the variation in OSCE results, with academic factors accounting for only 2%. Limitations of the study were that it was conducted in a single site and only addressed performance in the first three years of the course.

Taylor et al. (2005) conducted a study to examine the relationship between graduates' performance on a prototype of the National Board of Examiners' Step 2 CS (clinical skills) and other undergraduate measures with their residency directors' ratings of their performance as interns. The prototype CS exam is a 12 station OSCE, with questionnaires completed by standardized patients afterwards. GPA and prototype interpersonal score were the only significant predictors of intern performance. Variables correlated between .2 and .46, the highest being between GPA and quartile ranking. Limitations of the study are that the data is from one institution only and that residency directors' ratings as performance measures may be limited.

Basco et al. 's (2000) study in the USA, discussed earlier, showed no significant correlations between either academic profiles (grade point ratio and MCAT scores) or selection profiles from interview scores and third year OSCE scores. Eva et al. (2004b), discussed earlier, found that MMIs were a better predictor of performance on OSCEs than GPA and traditional interviews.

A study by Brothers and Wetherholt (2007), discussed earlier, examining the relationship between selection criteria for surgical residents and subsequent performance showed that medical school records (GPA) and USMLE scores correlated negatively with clinical performance. GPA correlations ranged from r=-.05 (knowledge) to r = -.26 (communication). Applicant personal characteristics and reference letters correlated most strongly with subsequent clinical performance. However, GPA and USMLE correlated positively with two other formal in-training tests, ABSITE and ABS QE. Residents who required remediation during training actually carried higher USMLE scores and GPA (both p<0.05).

Brown et al. 's (1993) study compared the records of twenty of 153 graduates from the class of 1983 at the University of California School of Medicine (USA) who failed to meet residency directors' expectations in Years One and/or Two with the records of twenty of their best-performing classmates in Years One and Two. The Medical School's admission data for these two groups of graduates was compared. The groups were quite similar in qualifications required for admission, overall academic achievement, and performance on standardized national examinations, with minor differences in performances on clinical clerkships. A questionnaire to residency directors asking them to rate (on a 5-point scale) 1982-1986 graduates' performance found that most of the poorly received graduates' problems during residency appear to have been personal and motivational. There were slight differences regarding academic strength, with three of the poorly received graduates and none of the best-received graduates receiving fewer than five (out of maximum twelve) points for GPAs, and two versus none with low MCAT scores.

Silver and Hodgson (1997) conducted a study to evaluate GPAs and MCAT scores as predictors of NBME 1 and clerkship performance based on students' data from one undergraduate institution (University of California, USA), so that grading system and curriculum would not be a confounding factor. Data for 88 students (out of 92 for the classes of 1990-1993) was available, and the sample was fairly heterogeneous in terms of uGPAs, MCAT scores, clinical performance grades and NBME 1 scores. Two regression analyses were performed to evaluate the relationship between clinical performances and data available at admission, and the relationship between these variables and NBME 1 scores. Results of the first regression analysis (F = 1.72, 5, 83, p>.1) showed that mean clinical performance was not related to any of the undergraduate predictors of performance (science or non-science uGPAs or MCAT scores). Results of the second regression analysis (F = 9.79, 5, 83, $p \leq .001$) showed that MCAT scores (ß = .29) and cumulative science uGPAs (ß = .25) were related to students' performances on the NBME1. The authors conclude that even when undergraduate grades are gained at the same institution, they are not useful in predicting clinical performance.

Taylor and Albo's (1993) overview of research carried out at the University of Utah School of Medicine (USA) in the 1960s and 1970s reports that physicians' performances in medical school, as measured by GPA, were almost completely independent of their later performances in their practices, regardless of specialty or years of experience. They concluded from their studies that 'academic achievement scores measure only a very narrow band of the extremely complex spectrum of skills and abilities used by physicians to practice medicine successfully' (Taylor and Albo, 1993, p.565).

### *7.4 Discussion*

Several studies have found evidence that academic grades predict future academic performance. Evans and Wen (2007) found that total undergraduate GPA was the most significant predictor in overall student academic performance in osteopathic medical education and MCAT scores were less predictive. Kulatunga-Moruzi and Norman (2002a) found that performance in licensing examinations was significantly correlated with GPA. Similarly, in later studies Eva et al. (2004b) found that GPA was the most consistent predictor on multiple choice exams measuring medical knowledge and Reiter et al. (2007) found that only the uGPA predicted progress scores and was statistically predictive of performance on multiple choice medical knowledge exams. Utzman et al. (2007) found that the odds of failing the USA National Physical Therapy Examination increased when GPAs decreased.

Some studies have shown evidence of longer-term effects of GPA. Cohen-Schotanus et al. (2006) found that a higher GPA of school-leaving examinations significantly enhanced study progress, with significantly less time to graduation, and was also associated with graduates qualifying in their preferred specialty and with higher scientific output. GPA scores had no effect on drop-out rate. McManus et al. (2003) found that A level grades had significant effect on time taken to gain membership qualifications, but did not predict diploma or higher academic qualifications or research publications. Peskun et al. (2007) found that residency rank in Internal Medicine was correlated significantly with undergraduate GPA in Internal Medicine, but not Family Medicine.

However, there have been mixed findings regarding GPA as a predictor of clinical performance. Hamdy et al. (2006) reported on eleven studies that had a combined low correlation between clerkship GPA and supervisor ratings, and on five studies of pre-clinical GPA as a predictor of supervisor ratings where the combined correlation was also low. The majority of papers discussed here did not show a correlation between previous academic performance and clinical performance. Silver and Hodgson's (1997) study showed that mean clinical performance was not related to any of the undergraduate predictors of performance (GPA and MCAT scores). Basco et al. (2000) showed no significant correlation between academic profiles (grade point ratio and MCAT scores) and third year OSCE scores and Eva et al. (2004b) found that MMIs were a better predictor of performance on OSCEs than GPA and traditional interviews. Brothers and Wetherholt (2007) found that medical school records (GPA) and USMLE scores correlated negatively with clinical performance. Taylor and Albo's (1993) overview of earlier research reports that undergraduate GPA was almost completely independent of physicians' later performances in their practices, regardless of specialty or years of experience. Two papers did report positive findings regarding future OSCE performance. Lumb and Vail (2004) found that A level grades and academic potential assessed on application were significant predictors of success in the third year OSCE. Taylor et al. (2005) found that GPA and prototype interpersonal score were the only significant predictors of intern performance on a prototype 12-station OSCE and residency directors' ratings.

### *7.5 Conclusion*

GPA has been found to be useful in predicting future academic performance; its utility in predicting clinical performance is less clear, with several studies showing no significant correlation. This raises questions regarding the weighting of grades in the admissions process, and which admission processes may best predict clinical performance.

## 8. Standardised tests

This section of the literature review focuses on the use of standardised tests as part of the selection process. Standardised tests are used to determine applicant's aptitude for the health profession (some examples of those used in the UK are summarised in the Table 1, below). Aptitude tests are designed to measure intellectual capabilities, general mental ability or intelligence. (McManus et al, 2005) this includes tests for thinking and reasoning; particularly logical and analytical reasoning abilities.

**Table 1: Aptitude tests currently used in the United Kingdom by medical school and courses (McManus et al., 2005)**

| | | |
|---|---|---|
| BMAT | Biomedical admissions test | Used by Cambridge, Imperial college, Oxford and University college London as well as three veterinary schools |
| GAMSAT | Graduate medical school admission test | Used for selection by Australian graduate medical schools. At present it is used by four graduate entry schools in United Kingdom. UK version GAMSAT:UK |
| MSAT | Medical school admission test | Used by three UK Medical school |
| MVAT | Medical and veterinary admission test | Developed in Cambridge and was a precursor to the biomedical admission test |
| OMAT | Oxford medical admission test | Developed in oxford and was a precursor to the biomedical admissions tests |
| PQA | Personal qualities assessment | Subtests of mental agility, interpersonal values, and interpersonal traits. Administrated in several UK medical schools on a research basis only. |
| TSA | Thinking skills assessment | Used by several Cambridge colleges in a range of disciplines, of which computer science is presently the predominant. |

### *8.1 Studies that report standardised tests as a predictor with later examination scores*

In Australia the Graduate Medical school admission test (GAMSAT) is designed to assess problem solving and data interpretation in the social, physical and biological sciences as well as critical thinking, reasoning and written communication (Elliott and Epstein, 2005). GAMSAT consists of three sections; GAMSAT 1: Reasoning in Humanities and Social Sciences, GAMSAT 2: Written Communication and GAMSAT 3: Reasoning in Biological and Physical Sciences and aims to assess the individual's capacity to undertake high level intellectual studies in a demanding course. It was introduced into Australia in 1996 and in the UK since 2000. It is designed to act as a filter in the selection process by providing objective data on defined areas of an individual's capability for medical study.

Coates (2008) conducted a cohort study to assess the criterion validity of the GAMSAT to predict performance in the first year of medical school. The three individual sections of the

GAMSAT and overall score were compared with GPA scores and interview measures. Graduate entry students (n = 351) across 6 medical schools in Australia were measured. The authors tested for concurrent validity assessing the relationship between GAMSAT, GPA and interview scores.

Findings from this study showed that GAMSAT section 3: had the strongest relationship with year 1 performance; possibly because both tests measure knowledge in the biological/physical sciences. A number of year 1 performance indicators did not correlate with components of GAMSAT and there negative relationships between year 1 performance and GAMSAT. The best linear model to predict year 1 performance included the overall GAMSAT score and GPA.

A major limitation of the study was the variability across the six institutions studied. This was reflected in the significant variability in responses rates across institutions and the problem with criteria of year 1 performance measures from different institutions. The authors did not include information on the interview measures. In some cases interviews reduced the predictive validity of the combination of selection methods while in others it improved the predictive validity. The authors did not include information on the interview measures used however interviews were shown to both add and reduce predictive validity in different schools which would suggest that there is a likelihood of a different interviews formats being used in the medical schools studied.

The Medical College Admission test (MCAT) was originally developed in 1928 and is currently on its fifth version. In the United States it is a standard requirement in many medical schools, and all schools approved by the American Osteopathic Association. The MCAT is composed of four subtests: biological sciences (biology MCAT) physical sciences (physical MCAT) verbal reasoning (verbal MCAT) and a written sample (written MCAT) and is usually taken in the year prior to medical school application.

The study by Hall et al. (1992), discussed above, examined the relationship between admission interview scores (both academic and non-academic criteria), MCAT, GPA and deans letter from students entering medical school. A significant relationship was found between MCAT scores and the dean's letter ratings (r =-.27, p = .04) with stronger dean's letter ratings associated with higher MCAT scores. MCAT offer useful predictive indications of how well someone will do in the pre-clinical curriculum. Similar conclusions were made by Silver and Hodgson (1997), discussed above, who evaluated GPAs and MCAT scores as predictors of NBME 1 and clerkship performance based on students' data from one undergraduate institution. Regression analyses showed that mean clinical performance was not related to any of the undergraduate predictors of performance but that MCAT scores and cumulative science were related to students' performances on the NBME1. The authors conclude that although MCAT scores are good indicators of NBME 1 performance, they are not useful in predicting clinical performance.

In the United States all dental schools require applicants to take the Dental Admission Test (DAT). The American Dental Association (ADA) administers the DAT which is designed to gauge general academic ability, comprehension of scientific information and perceptual ability. The DAT consists of which measures knowledge of natural sciences; biology (DAT-BIO) and general and organic chemistry (DAT-OC), reading comprehension (DAT-RC), quantitative analysis (DAT-QA) and perceptual ability (DAT-PAT). Other measures commonly used in the application process are undergraduate science GPA, general or cumulative GPA, letters of evaluation from faculty or members of the profession, relevant work experience in the field of dentistry, community service and interviews. Kingsley et al.

(2007) conducted a correlation study to examine the relationship between the. DAT, National Board Dental Exam Part (NBDE-1) and Dental School GPA (DS-GPA). The files of 225 students from three cohorts in a single US Dental school were examined. Elements of the DAT (DAT-BIO, DAT-RC and DAT-QA) were strongly associated with NBME-1 scores. DAT OC was not associated with either NBDE-1 or DS-GPA. The study was limited in using data based around only using three dental student classes in a single institution

The study by Poole et al. (2007), discussed above, examined the relationships between predictors, including a structured interview, and the clinical and academic performance of dental students in Canada. One of the predictors used in this study was the Canadian Dental Aptitude Test (DAT). DAT scores were related to some aspects of performance in the 1st and 2nd years, but not the 3rd and 4th year performance which have a more clinical practice orientation.

The predictive validity of a number of tests to determine the success of dental students in clinical and academic courses was explored by Chamberlain et al. (2005). Profiles of students (n = 89) at a Canadian dental school were compared with dental practitioners (n = 130) using the NEO-PI-R personality measure, a measure of professionalism (SPS) and cognitive ability using scores from Reading Comprehension Examination, the perceptual Motor ability Test and the Academic Average component of the DAT.

The DAT academic average scores correlated with both the DAT reading comprehension (r = .58, p<.01) and perceptual ability (r = .28, p<.05) components. Students who did well on Academic Average also did better on both Perceptual Ability and Reading Comprehension. Perceptual Ability and Reading comprehension were not significantly correlated with each other. Both Academic Average (r = 0.37, p<0.01) and Perceptual Ability (r = 0.36, p<.01) predicted performance in year one. Overall students who performed well in the first year of dental training received higher scores on the Academic Average and Perceptual Ability components of the DAT. Reading comprehension correlated with the third year clinical coursework (r = .42, p<.05); students who received higher reading comprehension scores performed better in the third year clinical coursework. Perceptual ability also correlated with scores of professionalism (r = .36, p<.01). Students who scored high on pre-admission Perceptual Ability component were more likely to receive favourable clinical performance assessments on the measure of professionalism.

### *8.2 Studies that do not report standardised tests as a predictors of either academic performance or clinical performance.*

Groves et al. (2007) looked at the relationship between medical students' scores on the GAMSAT and structured interviews, and their performance in medical school. Students (n = 189) volunteered in their second, third and forth year from Sydney and Queensland University Australia to participate. GAMSAT scores were compared with two other tests of Clinical reasoning problems (CRP) and diagnostic thinking inventory (DTI). No correlation was found between performance in GAMSAT and performance in the CRP. Weak negative correlation was found between GAMSAT and DTI (r = -0.31, p = 0.03). The correlation between GAMSAT and structured interviews was weakly negative for Queensland (r= -0.34, p<.01) and weakly positive for Sydney(r = 0.11, p<.01). The study found little evidence that GAMSAT and structured interviews are good predictors of performance in medical school.

The small sample size, statistical power of the study and low response rate (13%) gives the study a low confidence level. The use of a measure of clinical reasoning makes this study different from other studies involving the GAMSAT that have focused on academic performance. The authors conclude that their results highlight key questions on the use of entry tests as they have little predictive value of medical school performance.

The study by Evans and Wen (2007), discussed above, investigated the extent to which Medical College Admission Test (MCAT) sub scores predict the overall academic performance of osteopathic medical students. The study found that MCAT sub-scores had low predictive validity of overall academic performance (ß = -.01-.21) in comparison to undergraduate GPA. A similar conclusion was made by Peskun et al. (2007), discussed above. This study examined the predictive validity of academic and non-cognitive admissions components with medical school performance. The admission variables examined, including the MCAT, were compared with an OSCE score, Internal and Family Medicine clerkship grades and ward evaluations, and final medical school grade. There was no relationship found between MCAT and final application score and Family Medicine ranking only a slight relationship between MCAT and Internal Medicine ranking.

Hojat et al. (1993) conducted a prospective predictive validity study and examined a battery of nine tests; MCAT, psychosocial tests of general anxiety, text anxiety and locus of control, loneliness, neuroticism, sociability, stressful life events and perceived childhood relations. These tests were examined as to how well they predicted performance at medical school based on year one and two science exams and, clinical science exams and ratings of clinical competence.175 second year students (83% response rate) at Jefferson Medical College returned completed questionnaires. MCAT was shown to only predict the variance of 15% of basic science grades, 12% clinical science grades and only 4% of clinical competence ratings. This suggested that MCAT offered poor predictive validity for academic scores.

In contrast, psychosocial tests predicted 17% of the variance for basic sciences grades, 15% for clinical sciences and 14% for clinical competence. In combining these scores 32% of the variance could be explained for basic science grades, 25% in clinical sciences grades and 18% in clinical competence ratings. The authors observed that students who recorded fewer stressful life events, less anxiety, less loneliness, less externality in locus of control, and more sociability contributed significantly to predicting the ratings of clinical competence at medical school. The authors concluded that adding a battery of psychosocial tests to admissions criteria could improve the prediction of basic science grades, clinical science grades and clinical competence ratings in medical schools over and above what can be predicted from MCAT scores alone.

The study by Park et al. (2006), discussed above, evaluated the ability of pre-dental and pre-clinical benchmarks to predict clinical performance in a cohort of dental students at Harvard School of Dental Medicine, USA. The predictors were overall GPA, science GPA, dental admissions test (DAT) and performance on National Board dental examination. The outcome variables were measures of student performance and productivity during the clinical years (the four disciplines were: operative, major restorative and removable and fixed prosthodontics). The results of the study revealed a lack of any statistical associations between the predictors and the clinical outcomes.

### *8.3 Studies that report limitations of using standardised tests.*

The study by McManus et al. (2003), discussed above, compared measures of intelligence (ability); a standard intelligence test (AH5) on entry to Medical School, measuring verbal and reasoning ability (Part I “verbal”) and spatial ability (Part II “spatial”) and A-level (achievement). A level scores had a significant effect on time taken to gain membership qualifications, choosing to become a GP or leaving the register. In contrast the AH5 score which measures ability cannot independently predict membership qualifications or dropout. The study suggests that A-levels (achievement) rather than measures of intelligence (ability) are better predictors of performance.

The impact of practice bias and training is often raised as a general concern in relation to test taking. A study that looked at this specifically was Griffin et al. (2008) who examined the impact of coaching and re-testing on both the multiple mini interview (MMI) and the Undergraduate Medicine and Health sciences Admission Test(UMAT). A cohort of 287 applicants for entry in 2008 to the new School of Medicine at the University of Western Sydney participated.51% of the candidates interviewed had attended coaching prior to the test. In comparison to non coached candidates, coached candidates showed no difference on the UMAT scores for logical reasoning or understanding people but had slightly higher scores on the non-verbal reasoning part of the UMAT. However the difference was not significant after controlling for age, sex and the University Admission index. The findings would suggest that coaching might have a small positive effect on non-verbal reasoning component of the UMAT.

Fields et al. (2003) considered the gender effects on high stakes dental examinations which included the DAT. The scores from 451 dental students (128 females and 323 males) from six consecutive years at a US dental school were compared. Comparisons were examined between male and female scores on overall and science pre-dental GPA, DAT, National Board Dental Exam 1 and 2, North East Regional Board of Dental Examiners, and cumulative GPAs following second and fourth year. The study found men significantly outperformed equally prepared women in all areas with a percentage difference of between 2.29 to 8.01, except reading comprehension where women demonstrated a 3.25 percent difference and biology where they were comparable. It is highlighted however that although scores on the DAT do exist the magnitude of the differences are so slight that at time of postgraduate admission they become irrelevant, particularly as the passing, not the scores is the critical measure.

### *8.4 Discussion and conclusion*

There are a number of standardised tests currently used to recruit doctors internationally and in the UK. We have focused on the GAMSAT, MCAT and the DAT to consider the evidence of their ability to predict future performance.

The evidence for GAMSAT remains inconclusive. While Groves et al. (2007) found little evidence of predictive validity, Coates (2008) found some evidence that a model which incorporated GPA scores could predict year 1 performance. Notably these studies examined the relationship of the GAMSAT with academic performance not clinical performance.

The evidence of the predictive validity of MCAT remains limited to pre-clinical academic performance. Evans and Wen (2007), Silver and Hodgson (1997) and Peskun et al. (2007) all concluded little or no relationship found between MCAT and the studied clinical performance outputs. Hall et al. (1992) did find a significant relationship between MCAT scores and the dean’s letter ratings of student performance with stronger dean’s letter ratings associated with higher MCAT scores. Hojat et al. (1993) however reported more contradictory findings showing a low level of predictive variance which could be accentuated with the additional inclusion of a battery of non-cognitive tests. These studies do not offer any robust evidence toward the predictive validity of MCAT in academic settings and offer no evidence in association with clinical performance.

A number of studies have found relationship between the DAT overall score, or its components, and academic performance measures. (Kingsley et al.2007; Poole et al,2007; Chamberlain et al, 2005). However concerns over gender bias (Fields et al., 2003) and studies that found no associations (e.g. Park et al., 2006) contradict the overall efficacy of the DAT. Associations found are based on largely non-clinical performance, with the exception of Chamberlain et al. (2005). The findings by Poole et al. (2007) of a lack of association to performance at later stages on the course, where the curriculum included more work based clinical elements; question the predictive validity of clinical performance.

The literature as a whole that has examined the predictive validity of these tests have often focused on academic performance outcomes instead of considering the tests predictive validity of clinical performance outcomes. Beyond assessing academic aptitude there remains a scarcity of robust evidence in support of these tests and where focus has been on later stages of a medical course, where clinical content is higher, the predictive validity of the tests has often been poor. This suggests that caution should be given to generalising these findings to clinical practice and beyond an academic context.

## 9. Non-cognitive tests

There has been a clear focus within the literature of looking at the measurement of the non-cognitive elements of a candidate's application and the relationship this may have with future performance. Such approaches have considered the development of tests as a means to measure factors such as the big five model of personality factors, emotional intelligence, professionalism and interpersonal skills. A number of studies have attempted to establish the predictive validity of non-cognitive factors on performance, however limitations of the measurement tools and their sensitivity towards the contextual nature of the doctors role need consideration.

A body of studies refer to non-cognitive factors as representing aspects of the application that are not related to the cognitive scores gained through standardised tests. The non-cognitive factors in the literature encompass behaviours, attitudes, personality traits interpersonal skills and biographical data.

### *9.1 The basis for using non-cognitive selection approaches in selection*

An Australian study that considered the merits of solely using academic score in the selection process (Marley and Carmen, 1999) surveyed first and second year medical students, academic staff and school counsellors. Characteristics such as problem solving, critical thinking, communication skills, personal integrity, empathy skills and team membership were highlighted as necessary for being a doctor. The study emphasised selection approaches that move away from the singular use of matriculation scores to include qualities such as logical reasoning and critical thinking.

Halley (2008) aimed to identify the characteristics that general dentists preferred when hiring an associate dentist. The survey collected responses from 574 general dentists (14.8% response rate) in four US States of Idaho, Utah, Ohio and California. Respondents were asked to rank characteristics of a successful dental associate (from 1 most important to 5 least important). Highest ranking were: interacts well with patients (1.6) good relationships with principle dentist (2.4), interacts well with staff members (2.7), productive and efficient (3.4) and punctual (4.1). Factors that influence their choice of dental associate (1 most important to 8 least important) found that highest rank was personality (1.6), years of experience (3.7), completion of General Practice Residency/ Advanced Education in General Dentistry Programme (4.8), personal acquaintance (5.1), age (5.4), dental school attended (5.8), sex (6.5), military service (6.5) religion (7.6), other (8.8). The study concluded that interpersonal skills are regarded as the most important characteristics for a successful dental associate to possess. Personality was ranked as the factor that most influenced dentist's choice of dental associate. This list may not be exhaustive as the authors provided a prescribed list of characteristics, which is a limitation of the study, and other characteristics could have been generated but not previously considered.

In a study by Altmaier et al. (1989) key incidents from 31 senior radiologists across three medical schools were analysed through critical incident interviews. In describing each critical incident specific behaviours were elicited that distinguished doctors with good performance in the key incident and doctors with poor job performance in the key incident.187 critical incidents were identified and two physicians sorted them (inter-rater reliability of 0.89) into the six categories (knowledge, technical skills, attitudes toward self and conscientiousness, curiosity and conscientiousness and interpersonal skills. A Chi Square analysis was found to be non significant ($p<.05$) indicating the category distribution was similar both across sites and in comparison to a previous study. The pattern of distribution across the three locations in the study also did not differ greatly from the incidents obtained in the earlier study supporting generalisability and suggesting that it does not appear that one category is more important at one training programme than at another.

Across the critical incidents interpersonal skills (20% of incidents) and conscientiousness (43% of incidents) were the two categories of behaviour deemed as determining what was good and bad performance. Categories such as knowledge or technical skills did not feature as frequently in the incidents. The conclusion was not that categories such as knowledge are not important, but rather that doctors, following ongoing screening of cognitive factors throughout medical school, are a homogenous group in these categories. The study demonstrated consistency across setting in critical incidents. Behaviours that were more interpersonal in nature, or that were related to attitude, were cited as more of a requirement in a critical incident than behaviours involving technical skills and knowledge. For the purposes of selection this supports an approach involving the adoption of tools such as behavioural interviews and personality measures which measure these behaviours.

Non-cognitive factors are shown through the survey study by Marley and Carmen (1999) to tap into factors which are considered important characteristics for good doctors. This requires selection methods that go beyond the measurement of academic scores and introduce personality and behavioural concerns. Similar results were found by Halley (2008) when examining the factors that dental practitioners seek when selecting new associates. Within this different context non-cognitive factors were similarly reported as key characteristics that new dentists need to display. Taking a very different perspective Altmaier et al. (1989) examined the key features that distinguish good and poor doctor performance in critical incidents. The use of critical incident techniques enabled a focus on work performance and the findings, which identify conscientiousness and interpersonal skills to be most frequently reported. This study highlights an important link between performance and non-cognitive factors which should be considered a potential discriminator in the selection context.

### *9.2 Studies that report non-cognitive tests offer an alternative perspective on applicants*

Durning et al. (2005) examined the feasibility, reliability and validity of a supervisor evaluation form for first year residents as an outcome measure. A total of 1247 evaluations were collected for the 1559 graduates (80% response rate). Evaluation forms consisted of 18 separate items and an additional free text query. Factor analysis found that the evaluation form collapsed into two domains that accounted for 68% of the variance; professionalism and expertise. These two domains were compared to GPA and USMLE Step 1 and 2. Expertise correlated moderately with all three, whereas professionalism had no correlation with USMLE scores and only a weak correlation with GPA. The differences in these correlations support the notion of two distinct factors within the student evaluation. This study also demonstrates how tests of cognitive factors such as GPA and USMLE Step 1 and 2 measuring different constructs to professionalism, however assessing professionalism in selection is still as important as measuring expertise.

Carrothers et al. (2000) carried out a study to measure medical school applicants emotional intelligence (EI) using data from 147 applicants to a six year BS/MD program. Scores on an EI instrument which measured maturity, compassion, morality, sociability and calm disposition) were compared against the American College Test (ACT),a measure of knowledge and academic skills, high school GPA and a traditional interview assessment.

Low correlations were found between the EI measure and both the GPA and the ACT. Four of the dimensions of the EI measure (maturity, compassion, morality and sociability) had high correlations with the traditional interview. The interview did not contain any items measuring calm domain reflecting the low correlation found between interview and that domain. The study shows that a measure of emotional intelligence is a different construct to traditional cognitive methods and offers selectors different characteristics to be used in the selection of future doctors.

Collins et al. (1995), discussed above, studied the selection process going into a medical school in New Zealand and examined scores on interviews, a group exercise, a school report from the principal, and national exam scores. Candidates were interviewed twice and rated on seven non-cognitive attributes (communication, maturity, caring qualities/friendliness, awareness of community/political/social /medical issues, certainty of career choice, involvement in school activities, and involvement in community activities). In a group exercise candidates were rated on non-cognitive attributes (communication, leadership/organisational skills, friendliness, awareness of community issues and needs, knowledge of the nature of medicine as a profession, listening skills, and sensitivity and empathy to viewpoints and feelings of different individuals). School reports provided information on a candidate's communication skills, maturity, personal qualities, and involvement in sporting/cultural /community activities, as well as an overall rating of personal qualities. These three measures of non-cognitive attributes (interview, group exercise, and school report) were moderately correlated with each other (ranging from $r = 0.43$ to 0.62, $p<0.0001$), and were not correlated with a national exam score. These associations provide some evidence for convergent validity of using different measurement tools for non-cognitive characteristics and divergent validity as there was little association between the national exam and the non-cognitive measurements.

The studies by Carrothers et al. (2000) and Durning et al. (2005) showed the distinctiveness of a non-cognitive factor, such as emotional intelligence and professionalism, in a selection context and reported that their inclusion provides a different lens to look at the candidate from. A similar conclusion is taken by Collins et al. (1995) who further demonstrates that non-cognitive elements are distinct from traditional methods and offer a different perspective on which to base a decision upon.

### *9.3 Studies that report non-cognitive tests as a predictor of future performance*

The study by Hojat et al. (1993), discussed above, considered the contribution of non-cognitive measures in the prediction of medical school performance. Using a battery of psychosocial tests and MCAT the study examined how well they could predict performance at medical school. The MCAT was able to predict more of the variance in basic science and clinical science exams however the psychosocial tests predicted significantly more of the variance in basic science and clinical science in comparison to the MCAT alone. The authors also observed that students who recorded fewer stressful life events, less anxiety, less loneliness, less externality in locus of control, and more sociability were found to be rated more highly in clinical competence at medical school. The authors concluded that adding a batter of psychosocial tests to admissions criteria could improve the prediction of basic science grades, clinical science grades and clinical competence ratings in medical schools over and above what can be predicted from MCAT scores alone.

In the study by Frantsve et al. (2003), discussed above, the effects of personality traits on selection for an oral maxillofacial surgery residency programme in the US (n = 47) were examined. Personality was assessed by a standardised self-report personality measure

(Adjective Check List, ACL) and by interviewers on five personality traits (assertiveness, confidence, friendliness, motivation, and stress tolerance) during interviews. There were no significant correlations between the ACL and applicant ranking.

The study by Chamberlain et al. (2005), discussed above, examined the use of personality measures as predictors of the clinical and academic course success for dental students and compared their personality profiles with dental practitioners. Measures used included 3 components of the DAT, a measure of behaviours and professionalism (SPS) and the NEO-PI-R personality measure. The NEO PIR measure contains 240 items that were specific designed to assess the Big Five of Five Factor Model of personality (FFM) ;Extroversion, Conscientiousness, Openness to experience, Neuroticism and Agreeableness. A hierarchical regression was used to determine the relative contribution of the personality factors in predicting professionalism. Deliberation, a facet of Conscientiousness ($\beta = .33$, $p<.01$), and Ideas, a facet of Openness to experience ($\beta = .22$,$p<.05$) were significant predictors of professionalism.

The comparison of students and Dental practitioners found Dentists were higher in Agreeableness ($F = 8.04$, $p<.05$); they tended to be more compassionate. Dentists were lower in Neuroticism ($F = 7.59$, $p<.05$) they were more emotionally stable than dental students. Dentists were also higher in Conscientiousness, dentists were more organized, self disciplined and cautious ($F = 8.04$, $p<.05$).

Dental students were higher than dentists in Extroversion ($F = 23.46$, $p<.05$) students were more talkative and active than dentists reporting higher openness to experience scores (11.98,$p<.05$), dental students preferred novelty to familiarity and had more differentiated emotional states. Profile matching was also used to compare the overall student personality profiles to the mean profile for the dental practitioners and compared this with first year academic performance. It was found that students with personality profiles more similar to practitioners had higher first year grades ($r = .28$,$p<.05$). This study demonstrated that non-cognitive predictors, such as personality, have a role to play in the dental school selection process. Use of personality components, particularly conscientiousness and neuroticism, were good predictors of student's professional behaviour and first year performance.

The study by Poole et al. (2007), discussed above, examined the relationships between predictors, including a structured interview, and the clinical and academic performance of Canadian dental students. In examining the five-factor model of personality they found that conscientiousness was significantly related to many performance outcomes (from $r = .24$ to 0.47). The study suggests that conscientiousness shows potential as predictor of performance.

Similar findings in relation to academic performance were found by Chamorro-Premuzic et al. (2006) who compared the scores of 104 undergraduate psychology students from a British University on a battery of tests; five cognitive ability tests, the NEO-FFI personality tool that measure the FFM and atypical intellectual engagement scale. Neuroticism was significantly and negatively correlated with essay ($r = -.22$, $p<.05$) and exam ($r = -.25$, $p<.05$). Conscientiousness was significantly and positively correlated with essay ($r = .22$, $p<.05$), final project ($r = .20$,$P<.05$) and exam marks ($r = .31$,$p<.01$).

Studies, such as the three preceding studies, that are based on the FFM can attract two common criticisms; whether the FFM would apply equally well to non-volunteer samples, as is the case when respondents are job applicants, is there risk of socially desirable responding. Smith et al. (2001) examined three samples from a database of people who had completed a psychometric test; the Hogan Personality Inventory (HPI). A student sample (n = 2500), job applicant sample (n = 2500) and job incumbent (n = 2500) were used which differed along the social desirability continuum. The study found that the FFM was consistent across all of the samples and there was no evidence of a new 'ideal employee' emerging which would be the case if the social desirable criticism held true. The FFM also fit the job applicant sample better than the student sample which fails to support both of the critiques of the FFM. The study therefore concluded there was no evidence to suggest either the frame of reference or social desirability of FFM were valid criticisms.

Peng et al. (1995) assessed the usefulness of Cattell's 16 Personality Factor Questionnaire (16PF) in identifying personality variables in relation to the performance of medical students by considering the ability of the 16PF in differentiating problem students from non-problem students. The 16PF was designed to give a broad measure of personality that would be useful to practitioners in a wide range of settings. It uses185 items, and measures 16 personality factors with most items using a True/?/False format. The 16PF scores from Malaysian medical students (n = 101) were correlated with preclinical performance and problem student and non-problem student scores were compared. The study found that the personality factors of enthusiasm, venturesome, self opinionated, imaginative, experimenting, resourceful and driven were positively correlated with performance whereas being self – assured (placid or worrying) was negatively correlated. In addition students who had to re-sit their examinations in at least one module appeared to be more reserved, emotionally less stable and apprehensive (worrying) compared to those with straight passes. The authors noted that low levels of anxiety correlated positively with performance and that problem student had higher levels of anxiety or apprehension. The findings should consider the cultural context of the student studying in Malaysia; the findings distinguish how personality of the students may differ from western counterparts. The participants also all volunteered to the study which may also reflect certain personality qualities.

The benefit of the inclusion of personality tests within a selection approach was demonstrated by Hojat et al. (1993) as they were able to explain a greater proportion of performance variance than the MCAT. Higher performance linked to greater sociability and the experience of less stressful events suggest non-cognitive can describe a link to performance. Chamberlain et al. (2005) demonstrated facets of openness to experience and conscientiousness were predictors of performance while the students and dentists displayed different distinct personality profiles. The finding that higher performance was seen from students showing a closer 'fit' to the dentist profile suggests that candidate profiles far removed from this 'fit' might be likely to demonstrate below par performance in future roles. Conscientiousness being a potential predictor of performance was further supported by Poole et al (2007) and Chamorro-Premuzic et al. (2006) alongside a negative correlation with neuroticism. Peng et al. (1995) demonstrated in a Malaysian student sample personality factors that were related to higher academic performance and also underperformance. These studies demonstrate how personality can be predictive of future performance.

### *9.4 Studies that report non-cognitive tests as a predictor of poor performance*

Tamblyn et al. (2007) investigated the ability of clinical skills examinations (CSEs) to predict future medical complaints. A cohort of 3423 doctors who took the Medical Council of Canada clinical skills examination between 1993 and 1996 and who were licensed to practice in Ontario and Quebec participated. Members of the cohort had 1116 complaints filed against them with 696retained after the investigation. Of this cohort, 17.1% had retained 1 complaint.81.9% of the complaints were for communication or quality of care related issues.

A multivariate Poisson regression was used to estimate the relationship between complaint rate and scores on the clinical skills exam and traditional written examination. The patient-doctor communication score in the clinical skills exam remained significantly predictive of retained complaints (p<0.001) with scores in the bottom quartile explaining an additional 9.2% of complaints (95% CI, 4.7%-13.1%). The authors concluded communication and clinical decision making ability were important predictors of future complaints to regulatory authorities and current exams could be modified to test these attributes more effectively.

A study by Murden et al. (2004) examined whether observed professionalism deficiencies can predict poor performance in third year clerkships.44 students (6.7%) at Ohio Medical school were identified as having specific deficiencies. Deficiencies were identified through a physician and non-physician preceptor taking a doctor-patient relationship course that all students attended. Deficiencies were categorised into extreme shyness, poor process skills (unable to elicit, gather and organise relevant information) and negative attitudes.38 students were matched against a non deficiency group based on a 'MCAST' composite score of undergraduate science GPA, undergraduate college average MCAT scores and scores on the United States medical Licensure Examination (USMLE) Step 1. A significant difference was found between the deficient medical students and the matched controls on the third year clerkship performance (F = 7.45, p<0.05). The primary source of this difference was from those with the negative attitude deficiency; the inability to establish rapport with patients because of paternalistic behaviour, self centred behaviour or poor attitude towards doctor patient relationships. The authors concluded that those students who demonstrate deficiencies in professionalism and patient communication skills, particularly negative attitudes, in their first year of medical school tend to perform more poorly in third year clerkships than non-deficient controls. Some support to this finding of relating professionalism and attitude was provided by Heintze et al. (2004), discussed above. This study found professionalism in patient interaction was associated with Empathy and social competence as measured through the Affect Reading Scale(ARS). Positive relationships were found between empathy as measured by the ARS and tutor assessment of the student's social competence generally and of their interpersonal competence whilst being advised and of their contact with patients.

Knights and Kennedy (2006) investigated the effectiveness of selection interviews in deselecting students with negative personal characteristics. Students completed the Hogan Development Survey (HDS). This is a self-report measure of dysfunctional personality characteristics. The inventory is based on the diagnostic and statistical manual version 3 classification of personality disorders. Results showed that the HDS identified negative personality characteristics that were not detected in selection interviews. Principal component analysis (PCA) identified 4 factors which accounted overall for 67.3% of the variability in dysfunctional interpersonal behaviour. Factor 1: 'move away' from people, (moodiness, distrust, reluctance, independence) 58.7% had elevated scores, Factor 2: 'move against' people, (self-confidence, risk-taking, creativity) 73.6% had elevated scores, Factor 3: 'Diligent', (perfectionism). Scores were spread evenly across construct and Factor 4: 'Dutiful', (reliance upon others): 57% had elevated scores. In conclusion many negative characteristics associated with interpersonal problems are difficult to detect in an interview setting. The study shows that a significant proportion of new students report dysfunctional behaviours at levels that are worthy of attention by medical schools. The HDS could be used alongside selection interviews to identify dysfunctional tendencies.

These studies demonstrate an association between personality factors and underperformance. A similar association has been demonstrated by the UK National Clinical Assessment service which suggests that 29% of their referrals are (in the view of the manager) about behaviour alone (NPSA, 2006). The studies by Murden et al. (2004) and

Peng et al. (2004) show associations between behaviour and academic performance while Tamblyn et al. (2007) demonstrated a link between clinical decision making and communication with future medical complaints. Potential negative characteristics are evident at the selection stage however measures often used may not be sufficiently sensitive to identify them (Knight and Kennedy, 2006). Collectively these studies demonstrate a number of personality characteristics that are linked to underperformance but also highlight how more sophisticated selection methods, such as personality inventories, may be required in order to effectively detect them.

### *9.5 Studies that report contextual considerations in the use of non-cognitive tests*

Patterson et al. (2000) reviewed three studies that defined the competencies required for the job role of a GP. A mixed methodology of critical incident focus groups with GPs, behavioural coding of GP-patient consultations and critical incident interviews with patients were examined. The data collected enabled the development of a competency framework of 11 competencies with associated behavioural descriptors. These competencies were; empathy and sensitivity, communication skills, clinical knowledge and expertise, conceptual thinking and problem solving, personal attributes, personal organisation and administrative skills, professional integrity, coping with pressure, managing others and team involvement, legal ethical and political awareness and learning and personal development. The authors concluded that competencies required could be used to inform psychometric tests. The main limitation of the study was that the samples were only drawn from one area of the UK. In addition the authors highlight that the behaviours outlined in the results are not static and change as the job role changes and develops. The study highlights the importance of considering the behavioural attributes; not just academic and clinical skills in the selection process.

A related study (Patterson et al., 2008) applied the same methodology as Patterson et al. (2000) to three different Specialties; Anaesthesia, Paediatrics and Obstetrics and Gynaecology.14 competencies were identified; the 11 existing competencies found in Patterson et al. (2008) with the addition of two competencies; vigilance and situational awareness and teaching, managing others and team involvement were considered as two separate entities. This study also compared the competencies across specialties. Differences were found in terms of prioritisation both within and between the specialities studied. Differences were found across specialities for empathy and sensitivity, communication skills, organisation skills, professional integrity, team involvement and teaching. The similarities found across specialties could suggest that applicants should have a minimum standard across competencies however the differences identified also suggest that candidates should display an aptitude for the priorities of that Specialty.

Stratton et al. (2005) examined the relationship between emotional intelligence and empathy with clinical skills demonstrated in an objective structured clinical examination (OSCE).165 US medical students completed a 12 station OSCE following clerkship as a requirement to graduation. Within each encounter the students have 15 minutes to interact with a simulated patient (SP) and another five minutes to provide additional follow up information. Standardized patients then record each examinees performance using a predefined checklist – focusing on discrete, observable behaviours related to various aspects of the encounter such as taking a history or the physical examination and a three item checklist on communication skills. Emotional intelligence and empathy was measured through completion of the Trait Meta-Mood Scale (TMMS) that measures individual difference in recognising, discriminating and regulating emotions; and the Davis Interpersonal Reactivity Index (IRI) that measures dimensions of empathy. Results showed the TMMS and IRI subscales were significantly related suggesting overlap between EI and empathy. The subscales of Attention to feelings, Empathic Concern and Perspective Taking were significantly positively correlated with communication skills, while Empathic Concern and Perspective Taking were also negatively associated with physical examination skills. While this study highlights emotional intelligence, and empathy, as distinct from cognitive ability, it also raised questions regarding the use of EI in selection. This study demonstrates that the nature of the doctor's role may require a range of personality qualities depending on the situation and selection methods must be specific and contextually relevant.

A study by Kulatunga-Moruzi and Norman (2002) examined the validity of admission measures used to assess non-cognitive qualities for applicants to McMaster Medical school. The study compared three cohorts; those accepted to McMaster on the first round, those accepted on the second round and those who were rejected and accepted at another medical school. Admissions data analysed included GPA, simulated tutorial scores; group/communication skills/problem-exploration skills and personal interview scores), a written autobiographical submission and data for the Medical Council of Canada's Licensing Examination (LMCC) Part 1 and 2 were examined. GPA of those rejected by McMaster was significantly higher than those accepted in either first or second round. Those students offered places in the first round had significantly higher ratings on admissions scores than those rejected. Post hoc analysis of the simulated tutorial scores, group/communications skills and simulated tutorial problem-exploration skills, revealed similar findings. The mean scores for Personal interviews were significantly greater in those accepted in first and second round compared to those rejected. These findings demonstrate that the selection of McMaster students is heavily weighted toward non-cognitive measures. A comparison between LMCC performances found there was no difference between McMaster students and students who enrolled elsewhere.

The authors considered these findings of why McMaster students continue to perform well at LMCC studies but demonstrate lower GPA scores particularly when evidence demonstrate poor correlations between LMCC scores and non-cognitive factors (in comparison to GPA for example). They argue that the scores derived from measures of personal qualities, such as interviews and group exercises, are likely to not solely measure personality factors but a range of other extraneous variables. In contrast the strength of a measure such as GPA is that it is an average over time and variance due to extraneous factors are likely to have been cancelled out. Such measures often assume independence from context and a further argument presented is that these selection methods may be valid and reliable but performance in the context of admission data does not generalize to performance in the context of clinical examinations.

In examining a single construct, EI, across a number of different OSCE stations Stratton et al. (2005) highlight a concern regarding the assumption of core non-cognitive factors being context-free. The nature of the doctor's role is broad reaching and requires the student to be accomplished in displaying a range of appropriate behaviours; determined by the situational context. The study by Kulatunga-Moruzi and Norman (2002) highlights issues around the efficacy of measurement previously mentioned in the section. The distinction made between non-cognitive tests and GPA is an important one. The GPA is able to measure over time with extraneous factors cancelling each other out. For non-cognitive factors care must be taken to how far particular evidence is generalised across contexts and how much reliance is placed on a single measurement. The differences in the importance of competencies across specialties identified by Patterson et al. (2008) suggests that designed selection processes that incorporate non-cognitive elements need to consider contextual factors and have far these can be generalised to the different elements of the doctors role.

### *9.6 Discussion and conclusion*

The need for non-cognitive elements to be included within the selection process for doctors has been demonstrated through consultation on the desired characteristics of a doctors, and dentists, often focusing on non-cognitive rather than academic factors. A further argument presented is that as a whole the medical cohort are academically homogenous groups and therefore non-cognitive measurements offer a clearer means of discrimination. A review of critical incidents by Altmaier et al. (1989) provided evidence that non-cognitive factors can discriminate between good and bad performance during critical incidents. A number of studies have shown a relationship between personality and performance, as well as underperformance in both an academic context and in relation to medical complaints.

A number of personality factors have been found to relate to performance. Conscientious has been associated with good performance in critical incidents (Altmaier et al.,1989), a predictor of professionalism (Chamberlain et al., 2005), related to clinical and academic performance (Poole et al., 2007) and positively correlated with academic scores (Chamorro-Premuzic et al., 2006). Anxiety levels and high neuroticism have been associated with lower clinical competence at medical school (Hojat et al.,1993), correlated with negative performance and problems students behaviour (Peng et al., 1995) and negatively correlated with essay and exam marks (Charmoro-Premuzic et al., 2006). Good Patient to doctor communication skills were associated with good performance in critical incidents (Altmaier et al.,1989) empathy and social competence (Heintze et al., 2004) and lower likelihood of future complaints (Tamblyn et al., 2007). A number of associations have also been found between emotional intelligence constructs and performance. Empathic concern and perspective taking was negatively associated with physical examination and positively correlated with communication skills (Stratton et al., 2005). Empathy and sensitivity have also linked to professional competencies across a number of Specialties (Patterson et al.2000; 2008).

In general studies have suggested that non-cognitive tools should feature in selection and that they will add a different insight of the candidate than is currently available through traditional methods such as interviews or cognitive tests. Non-cognitive methods can be used to accentuate existing methods, e.g. Knights and Kennedy (2006), and may provide a more wide ranging insight to base a selection decision upon. The studies themselves refer to a range of personality tools and there is little consistency in practice therefore important considerations around the reliability, validity and utility must be taken in the usage of such measures.

## 10. 'White space'/personal statements

This section of the literature review focuses on the use of white space as part of the selection process. The definition we are using for the term 'white space' is the application process that involves the candidate writing a personal statement and how this was evaluated by the raters involved in processing applications. We have also included and other forms of white space that involve the candidate writing an essay or free text response which require candidates to provide evidence of thinking and analytic skills.

Written submissions, like interviews, are used to identify non-cognitive characteristics that are not identifiable from grade point averages or tests measuring knowledge. Assessments of non-cognitive elements are key to the selection process but achieving reliable and valid measurements of them is challenging. Glick (2000) states that there is universal agreement that outstanding cognitive qualities alone are not sufficient for medicine, however the essential personal qualities are even more difficult to evaluate.

There were very few studies that focused on 'white space', although a couple of studies included the personal statement as part of a batch of analyses they were using to examine the best tool for predicting performance. Crane and Ferraro (2000) and De Lisa et al. (1994) both reported that personal statements were among the items of evidence that were least important for ranking. A review article by Lee et al. (2008) reported not finding any specific studies correlating the personal statement with resident selection or future resident success.

### *10.1 Studies that highlight the potential strength of the personal statement*

The aim of a study by Sadler (2003) was to identify the differences between students who completed the nursing course and those who did not. The study was conducted at the School of Nursing, Michigan, USA. It examined the differences between 43 'non-completers' and 193 'completers' of the nursing baccalaureate. The study design was a cohort study involving retrospective comparison of data from those who completed the course ('completers') and those who did not. Measurements included comparing grade point averages, essays scores and a content analysis of the themes in the admission essay. The findings showed there were no differences between the grade point averages (GPA) of the completers compared to the non completers. The essay scores for each group showed minor difference, the scores for the non-completers were broader (1-20) compared to the completers (5-20). Essays were scored on organisation, focus, development of ideas, standard use of English and congruency on the values and norms of nursing. Inter-rater reliability for interview scores was moderately high (0.76).

The qualitative analysis identified two themes 'helping others' and 'family/personal care giving' that highlighted the main differences between those who completed and those who failed to complete the nursing course. The non-completers failed to describe a personal experience with a nurse or a family care giving situation. Nursing seemed to be external to them; nursing was something 'to do' rather than 'to be'. They did not personalise or internalise statements they made about nursing and the role of the nurse. This was in contrast to the completers who wrote about personal engagement with the care giving of family members and they developed personal goals from what they had observed. The main finding was that grade point averages and essay scores did not differentiate between those students who failed to complete the course from those who completed it. The main differentiation between the two groups was the internalisation of the nursing role. The limitations of this study are that it was conducted on one site and focused on nurses and the findings may not be transferable to other groups such as medicine. This study highlights the importance of understanding the clinical role early on: those who already had a good understanding of the role and internalised it were more likely to complete the training course.

The aim of a study by Amos (1996) was to determine whether medical school achievement is related to or predictive of performance on physical medicine and rehabilitation (PM&R) residency. Data was extracted from files of 205 former residents on the programme at University of Washington, USA. They found that entry into academic PM&R practice was predicted by an interest in such a practice in the personal statement of the residency application.

### *10.2 Studies that report on how personal statements are valued by raters*

The aim of a study by Travis et al. (1999) was to identify whether family practice residency directors, in the current pro-primary care climate will continue to prioritise the qualitative components of the application process over the quantitative (as they had 4 years earlier). Secondly, an aim was to determine if those residency programmes who are highly successful (high number of applicants) rank the components of the application process differently (value quantitative measures) compared to programmes that are less successful.

The method involved a national survey of all residency directors in the USA identified from the 1997 Directory of Family Practice Residency Programmes. Directors form the Accreditation Council for Graduate Medical Education approved residency programs in family practice were invited into the study. The response rate was 78%. The responses represented a variety of institutions and geographical regions. The main comparisons were carried out on the 260 programmes that provided data for both the 1994 and 1997 surveys. In both surveys the dean's letter, the personal statement were ranked first and second. In 1994 the transcripts (academic credentials) were ranked third and in 1997 this was the application form. There were no significant differences between those programmes that were rated successful compared to rated less successful, both rating quantitative measures in third and fourth positions after qualitative measures.

The personal statement was ranked first or second by all family practice residency directors, highlighting the importance of stated attitudes and aptitudes for family practice. This study highlights the value placed on personal statements above academic credentials in terms of selecting the person with the right attitude and philosophy for family practice. This study might be more relevant for selection at speciality level, but could highlight the general importance and value of the personal as reflected in qualitative measures over the impersonal as reflected in exam scores to ensure the right people are placed in the right job. The limitations of the study were that comparisons were focused on the responses from the same institutions in 1994 and 1997; the responses were not necessarily from the same individuals. Others members of faculty are also involved in selection; their views were not represented in the survey.

Galazka et al. (1994) conducted a survey of 282 residency directors from U. S. accredited non-military family practice programmes. They found that personal letters were rated as important by only 8% of the directors whilst interviews were rated as important by 51% and performances on clinical rotations by 36%. Selection decisions were highly influenced by candidates' performances in interviews and on clinical rotations.

### 10.3 Studies that highlight the weaknesses of the personal statement for selection purposes

The aim of Dirschl's (2002) study was to examine the inter-observer reliability of a scoring system designed to objectify the screening of orthopaedic residency applications. Forty resident applications were scored by six observers. The scoring system used included objective and subjective elements. The intra-class correlation coefficients for individual elements ranged from 0.28 to 0.98. The intra-class correlation coefficient was high for elements that were numeric but low for elements that were more subjective. Even with the use of careful definitions, raters had poor reliability in scoring elements such as letters of recommendation and personal statements.

The aim of a study by O'Neill et al. (2009) was to estimate the generalisabilty of the admission to medicine process during spring 2007. One of the measures examined included was a motivational measure based on a statement written in essay format. The assessment included five domains: written communication skills; knowledge of the chosen programme and profession; reflections on choice of programme and future employment plans. The written statement was prepared off-site and submitted with the application form. The five domains were assigned a plus or minus, used to guide an overall score of 0-100 on a global rating scale. The results of the motivational essay contained the largest proportion of undifferentiated error compared to the other admission variables which were qualification, knowledge and the interview. In addition to the undifferentiated error the motivation essay had a low applicant effect, a large rater effect and consequently the *G* coefficients were poor for the written motivation measure, but good for the admission interview. The authors acknowledge that competition for places is fierce and private coaching is advertised and model motivation essays have appeared in the internet. The authors reported that a sixth of the applicants had attended a coaching course and that their motivation essays were very similar, indicating potential ghost editing.

The aim of the Hanson et al. (2007) study was to compare the completion of an Autobiographical Screening tool (ABS) in controlled conditions on-site at the medical school interview, and off-site before interview via the web. The hypothesis being tested was that the controlled environment would lead to lower ABS scores, but that students would maintain their relative ranking. In addition it was hypothesised that a shorter time frame for completion of the questions would result in better differentiation between students. The study design was a paired comparison between off-site and on-site ABS completion. The study took place at McMaster University, Canada and involved 696 applicants selected for interview on the basis of GPA and off-site ABS. The ABS contained eight questions, six items were novel for on-site completion and two items were the same as off-site.

The results showed that the off-site scores were significantly higher than on-site scores and there were low correlations between the two scores. The items that were repeated were rated higher than the novel questions, but there was still a low correlation with off-site completions. Inter-rater reliability was better for off-site ABS. There were differences between raters, faculty assigned higher scores relative to the students and the community member's mean scores fell between the two. On-site ABS scores were moderately predictive of the multiple mini interview MMI (r = .30), but not the off-site scores. The disparity in the off-site and on-site ABS scores is of concern. The finding suggests that the difference could be due to deliberately fraudulent behaviour – i.e. help with completion. The validity problems make ABS unreliable and its use as screening tool very questionable for selection.

The aim of the McManus et al. (2005) study was to determine if the personal statement could predict later happiness or happiness with a medical career. The study included forty pairs of doctors, matched for sex. One member of each pair was known to be very satisfied and the other very dissatisfied with a medical career. Ninety-six assessors were used to judge which member of each pair was the happier. Of the 1,920 judgements made by the 96 judges 963 (50%) were correct, which was not significant. The study concluded that personal statement cannot validly be used by assessors to identify doctors who will subsequently be dissatisfied with a medical career. This study highlights that people can portray an ideal image to increase their chances of acceptance.

The aim of a study by Umansky et al. (2006) was to determine what factors predicted the ultimate fate of graduating plastic surgery residents. They analysed the data from 29 residents. Of particular note, they found no difference between academic graduates and non academic graduates with regard to their intentions in their letters of recommendation and personal statements. This study indicates that raters were not able to differentiate those more likely to have an academic career from those who were not. The particular limitation of this study was the small sample size.

### *10.4 Studies that have attempted to improve the reliability of free text submissions.*

Dore et al. 's (2006) study aimed to enhance the reliability and validity of an autobiographical submission (ABS) screening tool, used to screen pre interview non cognitive attributes. The study was conducted at DeGroote School of Medicine at McMaster University, Canada.3,907 applicants gave offsite ABS's, the top 1000 then had their ABS reviewed by three raters per applicant. The three ABS scores were combined with grade averages and the top 696 ranking applicants were interviewed. Interviews consisted of a 12 station MMI and an 8 question ABS.

Data from a subset of 30 randomly selected candidates was compared, first by rating the ABS on the traditional method (raters evaluated all ABS questions for a given candidate before moving onto the next candidate, vertical method) and the new method (where each rater scored all candidates on the first question before moving on to score the second question, horizontal method).

The scores for the ABS completed on-site were significantly higher than the ABS completed off-site. This finding was driven by the traditional scoring method where a higher mean score was given. The internal consistency (i.e. the average correlation between questions) varied with scoring method used. The traditional scoring method had greater internal consistency compared to the new horizontal scoring method. Inter-rater reliability was found to be higher for the ABS when completed on site and slightly higher for traditional vertical scoring compared to horizontal scoring (.81 vs .78). However, the ABS correlated better with the multiple mini interview (MMI) when the new horizontal scoring system was used (0.65 onsite).

This study highlighted that the new ABS scoring tool has potential as the preferred scoring system. The high internal consistency of the traditional vertical scoring system indicated that the ratings across questions were not independent and a halo effect may have introduced bias in scoring ABS. The new horizontal scoring system displayed greater inter-rater reliability. Compared to the MMI, the MMI has shown predictive validity (Eva et al., 2004b) as yet the ABS has not done so. The limitations were the small sample and that raters only received very minimal training.

### *10.5 Studies that show free text submissions can be more closely associated with clinical competence than basic science*

The aim of the Hojat et al. (2000) study was to examine consequential validity of the Writing Sample section of the Medical College Admissions Test (MCAT), whether score on the Writing Sample section of the MCAT predicts performance on clinical competence more than on basic science (based on the assumption that writing scores reflect analytic, organisation and problem-solving skills rather than knowledge base). The study was a between subjects design involving 1776 students at Jefferson Medical College Philadelphia between 1992 and 1999. The Writing Sample scores were classified into three groups: top, middle and bottom. ANOVA (and Duncan tests) were used for most comparisons and Kruskal Wallis for class rank.

Comparisons of the top, middle and bottom groups on the Writing Sample showed no significant difference for undergraduate science GPA or for the Biological or Physical science sections of MCAT. Significant differences were observed for undergraduate non-science GPA and the verbal reasoning test, clinical science and ratings of clinical competence and higher class rank with higher written sample. Verbal reasoning as a covariate reduced some, not all effects. Written sample scores were more closely associated with clinical competence than with basic science. The limitations of this study were that multiple analysis correction (Bonferroni) was not applied – although multiple significant results suggest alpha probability is not exceeded.

### *10.6 Discussion of strengths and weaknesses of using white space as part of selection*

White space offers one mechanism of access to the non-cognitive attributes that are an important part of selection. Identifying what are the important attributes and then determining how best to assess them is a vital first step. The Sadler study above indicates that white space can be used to assess the candidates understanding of the clinical role and potentially tap into the type of doctor the candidate aspires to, something which seems to be valued by selectors from certain specialties at least (Travis et al), although others prefer to rely on interview (Galazka et al.,1994).

One of the strengths of personal statements is that they can highlight a candidate's interest in a speciality and show that they have appropriate philosophy (Amos, 1996; Travis, 1990), although raters were not able to differentiate those more likely to have an academic career from those who did not, nor were they able to identify those satisfied with medicine as a career from those dissatisfied. Hojat et al.'s (2000) study highlights the potential strength of including essay submissions which seemed to be a better predictor of clinical performance than competence in basic science, on the assumption that it was drawing on analytic and organisation skill rather than knowledge.

The main weakness of white space has been the difficulty with measurement as highlighted in Dirschl (2002) and O'Neill et al. (2009) above. Improvements in reliability and validity can be made and Dore's (2006) study is one example. All of these studies emphasise the potential contribution white space has to selection. However, both the O'Neill and Hanson study warns of the potential for misuse when white space answers are generated off-site and answers are open to potential fraudulent assistance. Like Lee et al. (2008), this review also found no studies correlating the personal statement with resident selection or future resident success. In terms of cost-effectiveness this option is one of the least expensive compared to interviews and multiple mini interviews. Completion on-site would be more expensive than off-site but on-site completion would increase the validity. This option does not offer any particular merit for patient safety.

## 11. Structured records of achievement

A structured record of achievement is broadly defined as a document which captures and summarises the skills and achievements of a candidate. The UK higher education sector is moving towards this in the form of the Higher Education Achievement Record (HEAR; Universities UK 2007), which incorporates the European Diploma Supplement (The Europe Unit 2006) established by the 1999 Bologna Agreement between European nations as a description of higher education qualifications portable between member states. The motivation for this initiative is to aid the comparability of qualifications from different institutions (and countries), but also to provide greater detail to employers on the specific skills and experience candidates have, over that in a basic degree subject and classification. The record provides an extension of the degree transcript (less common in the UK than in the American higher education system), and provides a detailed breakdown of exams, courses and modules (with marks), as well as group work, and details of strengths and weaknesses in different areas.

The HEAR is only being piloted in the UK in the 2008-2009 academic year, so it is too early for its success to be evaluated. The review identified no papers explicitly using transcripts or similar records of achievement as part of a selection and admissions process. However studies which look at the grade point average, which is a single element of the transcript, are included elsewhere in this review.

The detail contained in the HEAR has some similarities with a learning portfolio, where a student or trainee assembles evidence of learning from different areas of their programme of study, although it is unclear how much primary evidence will be included in the HEAR. Undergraduate medical degrees already use portfolios (as indeed does postgraduate medical education), but as with other undergraduate assessments, these are not standardised, and vary in content and format between medical schools. The review found no use of portfolios in selection; while they are used for assessment, this tends to be formative and reflective in nature (e.g. Rees & Sheard, 2004; Driessen et al., 2005; Driessen et al., 2006).

There are studies which have looked at the structured scoring and analysis of achievements, although a formal record is not used. DaRosa and Folse (1991) evaluated a standardised student evaluation form for selection into residency. Forms were completed by the students' clerkship directors (n=137 students by 16 directors in 1986, n=76 students by 9 directors in 1987), and then similar forms completed by their residency programme directors at the end of their first year of residency. From these forms correlations were found between student and resident evaluations on six items in three areas (clinical abilities, knowledge and professional behaviour). There was more range restriction in 1987, clerkship directors only using upper half of scale. Moderate correlations between the clerkship and residency directors' ratings were found in 1987 (r=.45), with correlations on individual items between .43 and .57, compared to 1986 findings, where the overall correlation was not significant, and only two items correlated significantly (r=.32). NBME part 1 scores correlated with clerkship ratings (r=.30-.35) in both years, but residency evaluations just in the second year. Significant relationships were found between in-service examination percentile and some residency and clerkship ratings. Overall the study demonstrated feasibility and acceptance from those who completed the forms, but a low response rate of 52% suggests acceptance was not universal, and the low correlations throw doubt on its predictive validity.

Dirschl (2002) evaluated the inter-rater reliability of a scoring system applied to the applications of 40 randomly selected applications to an orthopaedics residency programme in the United States. Applications were reviewed and scored independently by six observers using a normalised 100-point scoring system applied to subjective and objective measures. Objective measures consisted of variables which may appear on a record of achievement such as the number of honour grades in basic science and clinical years of medical school, whether the applicant had achieved membership in Alpha Omega Alpha, the percentile score on the United States Medical Licensing Examination Part 1, the number of research projects done while in medical school, the number of abstracts or manuscripts published. The national ranking of applicants' medical school was also included. Subjective measures included determination of number of activities volunteered by the applicant that involved the use of gross motor and fine motor skills, the number of leadership and volunteer activities, an evaluation of three letters of recommendation, and evaluation of a personal statement. Overall inter-observer reliability (indicated by intra-class correlation) was high at 0.80, although below the .90 criterion desired for a reliable diagnostic test. Higher scores were recorded for numerical and objective measures and lower for subjective elements.

### *11.1 References and letters of recommendation*

Personal references have long been part of selection, but these days are often fairly structured in content – in part because of validity issues, and in part to avoid litigation. Several US studies have looked at structured letters of recommendation, formal 'Dean's Letters', or similar reports which effectively provide a systematic overview of a student's performance which will contain much of the information in a structured record of achievement.

Girzadas et al. (2004) examined a standardised letter of recommendation (SLOR) for selection into a US emergency medicine residency, and whether outcome related to the genders of the applicant and author of the letter. The emergency medicine SLOR uses a checkbox format that focuses on a number of attributes: work ethic, willingness to assume responsibility, ability to develop a treatment plan, ability to work with others, ability to communicate a caring nature to patients, how much guidance the applicant will need in the residency and a prediction of the success of the applicant. The author also has to grade the applicant overall as 'outstanding', 'excellent', 'very good' or 'good', and rank them relative to other applicants for whom they have written letters.835 SLORs were analysed, and a relationship between author and applicant gender, and application outcome was found. The female applicant/female author combination was twice as likely to result in a 'guaranteed match' recommendation than any other gender combination (odds ratio=2.0). A limitation of the study however was that the amount of contact between author and applicant was not controlled for, and so may have been a confounding factor.

Other studies of letters of recommendation have found conflicting results. Dirschl et al. (2002) found that ratings of letters of recommendation for applicants to a US orthopaedic residency programme did not correlate with later performance on in-training examinations, board examinations or faculty ratings of residents. Hayden et al. (2005) found that a letter of recommendation does predict performance on measures of clinical but not academic success, although only for graduates of lower ranked medical schools.

Lurie (2007) compared the ranking in a dean's letter with programme directors' ratings of trainees, 9 months into internship. The study used a quasi-experimental design, with the dean's letter ranking ('good', 'very good', 'excellent' and 'outstanding') treated as a four-level independent variable on which data from residency programme directors was compared.

Residents were rated by programme directors on a 13 item questionnaire which was treated as a single factor (Cronbach's alpha=.94). Directors were also asked if respondents agreed with the classification in the dean's letter. Analysis of variance found differences between the top two dean's letter classifications and the remainder. Programme directors disagreed with the dean's letter ranking for only 19 of 104 students, with 14 referring to 'Good' (the bottom group), 7 feeling it overestimated trainee and 7 that it underestimated. The remainder were in high categories and mostly felt the letter underestimated the student. There was more variability between programme directors' ratings of the 'good' and 'very good' graduates. The study highlights the potential use of references to indicate strong performers, but a lack of predictivity for average performers.

Naylor et al. (2008) looked at the relationship between a range of experience-related variables (including class rank, clerkship grade, research experience, publications) as well as dean's letter and interview score on performance on American Board of Surgery examination, in-training assessments and residency programme directors' evaluations. A lack of superlatives in the dean's letter was found to be a predictor of an unsatisfactory outcome.

### *11.2 Discussion and conclusions*

Evidence which may be contained in a structured record of achievement has some predictive validity, and the standardisation of such information may be beneficial. The UK higher education sector is moving towards a universal record, and medical education may be required to co-ordinate with this (for the purposes of intercalated degrees and the benefit of students who may transfer away from medicine).

## 12. Assessment centres, work samples and simulations

### *12.1 Assessment centres*

This section describes studies which have looked at novel methods of ascertaining behavioural or interpersonal skills. The methods discussed in these studies are those which are used as a direct part of selection, and do not include OSCEs and similar assessment-oriented behavioural examinations. These tasks are often completed as part of 'assessment centres' in which candidates may complete a battery of tasks in one day, including interviews, cognitive and non-cognitive tests which are discussed in other sections of this review. Assessment centres consist of "a standardised evaluation of behaviour based on multiple inputs" (International Task Force on Assessment Center Guidelines, 2000, p.2). They are similar in some ways to am OSCE or multiple-mini interview, but will contain stations which may each look at a number of cognitive or non-cognitive elements, rather than each addressing only one. Assessment centres also have a longer history outside medical education, dating back to the Second World War (Patterson & Ferguson, 2007).

Empirical findings on the predictive validity of assessment centres in non-medical fields have been promising. In a widely cited meta-analytic review, Gaugler et al. (1987) reported a corrected mean validity of .37 for assessment centres. Gaugler et al. (1987) also examined the predictive validity for five different criteria: overall job performance (.36), management potential (.53), performance on the dimensions rated in the assessment centre (.33), training performance (.35), and career advancement (.36).

Arthur et al. (2003) examined the predictive validity of common assessment centre competencies. The competency demonstrating the greatest validity was problem-solving (.39), followed by influencing others (.38), organising and planning (.37), communication (.33), drive (.31), and lastly, consideration/awareness of others (.25). Arthur et al. (2003) also reported that, together, the competencies produced a multiple correlation of .45 ($R^2$=.20). Further regression analyses identified a subset of four competencies that accounted for the most unique variance in assessment centre scores: problem solving, influencing others, organising and planning, and communication. Neither drive nor consideration/awareness of others made a significant contribution to the model when the other four competencies were included.

Patterson et al. (2005) looked at the predictive validity of an assessment centre comprising five components on six dimensions of practice. The assessment centre consisted of a series of simulation exercises observed by raters (details are not provided). Looking at data from 46 GP registrars who had passed through an assessment centre, and 20 who had not, they found that there were few significant differences between the two groups' supervisors' reports after three months' practice (only on problem solving and clinical expertise). Within the assessment centre group however, those who had scored highly did tend to score more highly in their supervisors reports on most dimensions.

Randall et al. (2006) describe an assessment centre implemented for selection into paediatric training in a UK deanery, and the relationship between task and interview performance. The assessment centre consisted of a structured interview, a simulated consultation with a parent of a young child, a group exercise discussing the prioritisation of tasks (clinical and non-clinical), and a reflective written exercise based on the group task. Scores gained on the structured interview were significantly correlated with those on the group exercise and the simulated consultation (r > .4). The written exercise correlated with the group exercise (r > .5). However the correlations were low between the simulated consultation and written exercise (r=.21) and group exercise (r=.12). The non-interview tasks had some effect on outcomes, with three of 27 candidates not offered places who would have been on interview alone.

Candidates' perceptions of the assessment centre were also gathered. The vast majority indicated that the assessment centre gave them more opportunity to demonstrate their abilities than other medical selection processes experienced and that the content was more relevant to work in paediatrics. The study had a relatively small sample (n=27), drawn only from one area of the UK, although Randall et al. (2006) found similar results from sixteen participants in a pilot assessment centre for obstetrics and gynaecology.

An example of an assessment centre described by Ziv et al. (2008) involved a standardised-patient simulation task for the evaluation of personal and interpersonal skills as part of recruitment to an Israeli medical school (n=283 and 280 across two years). The task consisted of eight stations rated by standardised patients and faculty members: three scenarios with standardised patients, two short interviews about the scenarios, a standardised personal interview concentrating on attitudes towards the medical profession, and two group sessions where candidates worked in groups of six on problems such as ranking statements relating to the medical profession. Correlations with cognitive test scores were found to be low, indicating non-cognitive elements were being assessed by the stations. The system was found to be reliable in terms of inter-rater reliability and test-retest consistency.

### *12.2 Simulation and work sample*

The simulation tasks included in assessment centres may constitute a form of 'work sample'. This term refers to measures or assessments which provide a sampling of the candidate's workplace behaviour. Two studies of applicants to a Dutch medical school describe a method of selection termed by Oosterveld and Cate (2004) a 'study sample assessment procedure' (SSAP), which is a form of work sampling assessing both information processing and communication skills. Candidates are put into pairs, and each spends an hour reading a 3-5 page text about a different disease. They then explain what they have learnt to their partner, followed by two 15 minute interviews with a standardised patient about both diseases (that they studied, and the one their partner studied). The method studies knowledge retention and understanding, communication with peers, and communication with patients. A panel of three views these interviews from behind one-way glass, and rates them on standardised forms.

Cate and Smal (2002) looked at the content of the interviews in terms of the quality of information given to the patients, the attitudes expressed and the communication skills demonstrated (n=61). While the results are not reported in depth in this paper, reliability coefficients of between .6 and .92 are quoted. The study reported by Oosterveld and Cate (2004) compared the reliability of the SSAP with a structured interview by a different panel of three, and free text responses to an application form (n=172). It found that the SSAP showed comparable reliability to the interview, and to the overall process including all three elements (with G-coefficients greater than 0.7), and great than the application form alone.

Another example of a work sample method was described by Kogan and Shea (2005). This was a simulated case write-up tool as an assessment method. Standardised forms were rated by multiple raters, with 14 specific qualities of the write-up and a global measure. Data from 165 US medical students found that the tool was a reliable method of capturing skills for clinical writing skills, and may have had educational value in improving those skills. While reported as an assessment method, this may indicate a means of identifying skills during selection.

Bland et al. (2005) looked at the feasibility and validity of different approaches to scoring of a script concordance test for assessment or selection. A script concordance test may provide evidence of clinical reasoning skills and knowledge base as part of an assessment centre, although it is usually a summative assessment method. Candidates indicate their perception of the importance of different information or interventions, and their responses are compared with those of an expert group. Kreiter et al. compared responses of 85 respondents of different grades (from medical student to expert) to an expert panel. Outcomes using either direct match to mean expert scores, deviation from mean scores, or deviation from modal expert scores were compared. Three and five-point scales were also compared. The scores varied with the expertise of the respondents, indicating construct validity, and it was found that a system which took modal expert responses rather than means resulted in greater reliability.

Chamberlain and Searle (2005) reported an evaluation of a pilot teamwork assessment task in which applicants to a UK medical school were observed discussing a non-clinical ethical issue in small groups. The intention was to examine aptitude for problem-based learning type interactions, but the method may be generalisable to other course types. Behaviours which were identified as desirable were rated on a four point scale, and summed to produce an overall score. Low concurrent validity was found in comparing scores with selection interviews (r=-.32), although there was overlap between the individuals clustered at the top and bottom of interview and teamwork assessment rankings. Eleven out of 44 candidates who passed the interview scored below pass standard in test, while 23 of the 69 candidates who passed the teamwork assessment failed the interview. This could indicate a lack of concurrent validity, or that the assessment is measuring a complementary construct. No predictive validity was observed in terms of undergraduate year one assessments, but the sample for whom such data was available was very small (n=16). Problems of the subjectivity of interpreting behaviour as belonging to a given desirable category, and of potential halo effects leading to mis-identified behaviour. The method does have apparent potential however.

Lievens et al. (2005) described the use of a ‘situational judgment test’ (SJT) which aims to capture the skills of interpersonal judgment which would be required in a clinical setting, although not on a clinical task. In the study 7197 applicants to Belgian medical schools across four years completed a test in which they viewed 30 video scenarios, responding immediately after each one to a multiple choice question about interpersonal issues raised. Scores were compared with the cognitive test used as part of the centralised admissions process for Belgian medical schools, and a low correlation was found (r=.19). The SJT was found to explain variance in grade point average (GPA) across the four year degree beyond that explained by the cognitive test, with greater GPA variance explained if the course had a higher interpersonal skills component. The conclusion was that the SJT did have predictive validity along a dimension distinct from the cognitive test. The implication is that interpersonal skills are relatively static – the range does not compress by the end of the course.

For admission to specialty training/residency programmes, higher fidelity simulation may be appropriate. Savoldelli et al. (2006) looked at the performance of Canadian anaesthesia residents on a simulator, compared to their performance in an oral examination immediately preceding it. Both methods compared performance on resuscitation and trauma scenarios. Scores were consistently higher on the simulator, with higher pass rates and fairly low correlations with the interview scores (r=.52), but there was agreement on those who failed in each method. The majority of variance was found to come from participants, regardless of scenario or modalities. The conclusion was that the simulation picked up participant qualities different from those detected in interviews alone. Problems uncovered in simulation were procedural rather than knowledge-based: a resident who 'knows how' may not be able to 'show how'. Both methods had comparable ability to filter failing trainees. A questionnaire found the majority of candidates felt positively about simulation as a useful part of assessment. While the study looked at an assessment context, the findings may transfer to specialty recruitment, with the simulator providing a demonstration of practical skills.

### *12.3 Discussion and conclusions*

This section has discussed studies which present evidence relating to assessment centres, and on simulation and work study tasks which may be most effectively situated in such centres. Assessment centres may have value in allowing a range of cognitive and non-cognitive elements to be evaluated in a time-efficient way, although they do require organisation and resources. Their real value though will depend on the stations or tasks which are incorporated. Work sample and simulation approaches can provide useful ways of accessing complex skills and aptitudes.

An assessment centre is not a singular off-the-shelf method of selection, a format for the administration and delivery of selection. Once appropriate methods of selection have been chosen, whether interviews, simulations, practical tasks or teamwork exercises, it may provide an efficient way – for both applicant and recruiting organisation – of running the selection process.

## 13. Discussion

A need has been identified to review the process of selection and recruitment to the first year of the Foundation Programme. Perceived weaknesses in the existing process have led to concerns that the recruitment of new doctors is lacking in robustness, validity and reliability. Two stages in developing a new or revised method are identified: defining the purpose and requirements of selection, and then establishing the appropriate methods of selection (Guion, 1998; Patterson & Ferguson, 2007).

### *13.1 Purpose of selection*

In reviewing selection procedures, it is important to consider the reason for selection – what are candidates being selected for? The papers discussed in this review have addressed three main areas of selection – undergraduate, initial postgraduate training, and advanced postgraduate training (specialty programmes). Of the three, the second has most in common with selection to Foundation Programme, and appears to be the most under-studied. Selection of medical students must by definition focus on aptitude and potential for clinical work as the majority of applicants will have no clinical experience. Predictors derived from performance in non-clinical domains, while that for specialist training can be based much more on clinical experience, and performance in directly relevant examinations. Entry to Foundation Programme level training must make inferences from little clinical practice, as medical students do not have the opportunity to take real clinical responsibility, and so must be based more on academic and non-academic analogues and aptitudes.

To ensure the right dimensions are measured in selection a competency framework should be identified based on a job analysis. Appropriate selection methods can then be chosen to assess these competencies (see Randall et al., 2006; Patterson et al., 2008). The principle is akin to that of blueprinting assessments to a curriculum, as used in the development of the Foundation Programme assessment tools (e.g. Archer et al., 2008). For Foundation Programme year 1 however, a detailed job analysis, based on observation, may be hard to carry out and less useful in practice because the role of the Foundation Year 1 doctor varies so much between placements – locations and specialties – within the course of a year. A job analysis may however be carried out by reviewing the General Medical Council policy documents which define the competencies required of a Foundation doctor. *Tomorrow's Doctors* (GMC 2003, although currently being revised) defines the competencies which a medical student must have at graduation, and so effectively specifies the minimum standards required of a new doctor. *The New Doctor* (GMC 2007) sets out the standards which must be achieved during Foundation Programme, and so indicates the aptitudes which applicants should be able to demonstrate, even if their experience at application does not provide direct evidence. However it should be noted that both of these documents are focused on education, and the competencies which are of importance to employers during selection may have a different, more service-oriented, focus. However, the key elements are the same – that the doctor should be able to perform a range of clinical tasks professionally and safely.

### *13.2 Methods of selection*

This review has considered a breadth of evidence concerning various selection methods. The main strengths and weakness of each approach are summarised in Table 2.

Different methods address different aspects of an applicant, and the literature identified different methods for different dimensions of performance – often simplified into 'cognitive' and 'non-cognitive', but with varying levels of complexity. Some methods attempt to measure or otherwise capture these dimensions directly, while others such as personality testing capture variables which may determine behaviour in applied domains. The former aims may use established summative examinations as indicators of expected future performance, while the latter may use tools solely for selection. There are also differences in the level of performance on Miller's pyramid (Miller, 1990) that the evidence relates to, namely whether it is related to knowledge, behaviour, or actual practice. For indirect measurement, indications of aptitude may come from expressed attitudes, personality dimensions, or psychomotor skills.

The main findings are that cognitive measures (such as national exams, grade point averages and standardised tests) can assess knowledge and are moderately predictive of later cognitive measures such as membership examinations. However, cognitive tests have low predictive validity for clinical practice which involves skill-based qualities. It has been stated that there is universal agreement that outstanding cognitive qualities alone are not sufficient for medicine (Glick, 2000).

Adding more cognitive tests at a national level would not necessarily add value then, in terms of ascertaining an applicant's knowledge over and above the standard demonstrated by completion of a medical degree. It would also risk losing the diverse ecosystem of medical curricula, which provides welcome heterogeneity in the workforce. The ranking of graduates which would likely follow from such an exam may also unfairly disadvantage some doctors – a normative distribution can exaggerate differences which may be small on any meaningful criterion. An exam is also a limited snapshot of cognitive ability, compared to the multiple sampling that a longitudinal measure such as the grade point average provides.

The review has identified the importance of addressing non-cognitive components during selection. The recent study by Illing et al. (2008) on the preparedness of medical graduates for practice highlighted that the main area of weakness when starting Foundation Programme was lack of exposure to clinical practice. Evidence from this review indicates that additional components of selection should access both evidence of clinical practice gained as an undergraduate, and the non-cognitive qualities which will enable good practice.

Evidence of clinical experience may be best gained through retrospective evidence from medical school in the form of a structured record of achievement, or additional assessments for the purpose of selection for the benefit of employers. These assessments may be conducted in the framework of an assessment centre which may contain simulations of practice, including simulated consultations and work samples. A similar method which provides opportunities for the assessment of both clinical practice and non-cognitive attributes is the multiple mini interview, which has a growing body of evidence of its predictive validity. The methods which the assessment centre/multiple mini interview structures allow for enable a greater assessment of the 'shows how' and 'does' elements of Miller's pyramid (Miller, 1990) of clinical competence, compared to cognitive assessments which can only demonstrate the 'knows' and 'knows how'.

Other non-cognitive tests measuring elements such as personality and behaviours can provide additional indicators of future performance, but may be divorced from actual clinical performance. That said, relatively stable attributes such as conscientiousness have been shown to relate to performance.

Selection processes for the Foundation Programme must take account of the fact that the population of applicants will cover a narrow range, and all but a very few will be competent to perform their job and develop as professionals. Selection is therefore not about sorting 'wheat from chaff', or people who will be good doctors from those who will not, but rather making sure that applicants who are the 'best fit' for an employer are selected, and that any who do 'fit' but have particular areas for development are identified and can be supported.

Selection methods must also satisfy their stakeholder constituencies. Employers want to retain their independence, and exercise choice to be confident that their workforce is able to deliver the service and maintain patient safety. Students must feel that they are able to present their capabilities to their best effect, and are not being disadvantaged by methods which do not effectively assess their competencies. Employers and applicants need to feel they have appropriate control over the outcome to feel that the process is fair. Interviews satisfy both candidate and employer in terms of perceptions of control. The traditional, unstructured interview does not however have the necessary validity or reliability to measure appropriate qualities and competencies, and is open to interviewer bias. Structured interviews have higher validity and reliability, and are typically preferred over traditional interviews. The multiple mini interview may add further robustness, while satisfying stakeholder preferences.

### *13.3 Conclusions*

The main findings of this review are that cognitive tests are moderately predictive of later cognitive tests. However, non-cognitive elements need to be considered to ensure the doctor is able to perform well. The important non-cognitive elements are clinical practice and personal attributes. An assessment centre approach provides a framework allowing for a range of methods to be used, which may include multiple mini interviews, personality tests and skill-based assessments.

**Table 2: Strengths and weaknesses of selection options**

| Strengths | Weaknesses |
|---|---|
| **Interviews overall**<br>• Positively regarded by candidates and interviewers<br>• Provides opportunity to assess cognitive aspects<br>**Structured interviews**<br>• Less subjective<br>• Reduces risk of bias<br>• More defensible against legal challenge<br>• More reliable and valid | • Can be costly and resource intensive<br>**Structured Interviews**<br>• More evidence required for predictive validity in the medical field<br>**Unstructured interviews**<br>• More subjective<br>• Increased risk of biases<br>• Open to risk of legal challenge<br>• Less able to identify risk to patient safety<br>• Less reliable and valid |
| **Multiple Mini Interviews**<br>• Candidate and interviewer reaction positive<br>• Assessment of multiple non-cognitive attributes is context specific<br>• No benefit from coaching, and test violations should not affect score<br>• Some evidence of ability to predict clinical performance<br>• Scenarios could include patient safety issues<br>• More defensible against legal challenge<br>• Reduces bias<br>• Reliable and valid measure of non-cognitive attributes<br>• Identify attributes to match competency framework<br>• Tailored to required tasks e.g. ethical reasoning, empathy | • Labour intensive to develop<br>• More evidence on predictive validity needed for postgraduate use |
| **National exams**<br>• Measures cognitive ability<br>• Moderately predictive of cognitive tests administered later<br>• Standardised<br>• Objective<br>• OSCEs can be used to assess clinical performance<br>• More defensible against legal challenge | • Low predictive validity for clinical practice<br>• May lead to ranking of medical schools<br>• Knowledge test does not translate to assessment of patient safety |

| Strengths | Weaknesses |
|---|---|
| **Undergraduate Grade Point Average**<br>• Measures cognitive ability<br>• Moderately predictive of cognitive tests administered later<br>• Standardised<br>• Objective<br>• OSCEs can be used to assess clinical performance<br>• May have better reliability than a one off national exam or test as GPA reflects combination of more data<br>• More defensible against legal challenge | • Low predictive validity for clinical practice<br>• Knowledge test does not translate to the assessment of patient safety<br>• Medical Schools have different standards and different exit tests |
| **Standardised tests**<br>• Measures cognitive ability<br>• Moderately predictive of later cognitive tests<br>• Standardised<br>• Objective<br>• More defensible against legal challenge | • Currently only used at selection for medical school<br>• Low predictive validity for clinical practice<br>• Knowledge test does not translate to patient safety |
| **Non-cognitive tests**<br>• Offer a clearer means of discrimination within an academically homogeneous group<br>• Certain personality traits seem to be linked more to performance eg conscientiousness<br>• And underperformance eg communication problems | • Lack of evidence on best tool to select<br>• Lack of evidence to confirm potential impact<br>• Range of tools may need to be validated on medical graduates<br>• Non-cognitive tests do not assess risk to patient safety–although some underperformance attributes might be indicators |
| **White space**<br>• Can highlight individual interests<br>• Can highlight understanding of future role<br>• On-site completion not open to ghost writing | • Poor reliability in scoring – unless specific system set up<br>• Off-site completion is open to ghost writing<br>• Model answers on internet<br>• Lack of evidence for predicting future performance<br>• Personal statements do not assess risk to patient safety<br>• Less defensible against legal challenge due to subjectivity |
| **Assessment centres**<br>• Candidates have multiple opportunities to demonstrate their abilities<br>• Content of assessments can to work focused<br>• Can demonstrate knowledge of both practical and clinical skills<br>• Scenarios could include patient safety issues | • Need to ensure the appropriate assessments are used for F1 selection<br>• Relatively new method of selection in the medical field – no competency framework developed<br>• Could be relatively more expensive than other selection techniques |

| Strengths | Weaknesses |
|---|---|
| **Structured record of achievement**<br>• Has the potential to add to GPA record by providing a record of experience in clinical practice<br>• Would be stronger if the record was standardised and national<br>• Record could include patient safety issues | • Lack of existing evidence<br>• Would be weaker if record was local and not standardised<br>• Record would need input from tutors/clinicians increasing cost |

## 14. References

Alexander, G.L., Davis, W.K., Yan, A.C. & Fantone, J.C., 2000. Following medical school graduates into practice: residency directors' assessments after the first year of residency. *Academic Medicine*. 75, pS1517.

Altmaier, E. et al., 1989. Cross-institutional stability of behavioral criteria desirable for success in radiology residency. *Investigative Radiology*, 24(3), pp.249-51.

Altmaier, E.M., Smith, W.L., O'Halloran, C.M. & Franken, Jr. E.A., 1992. The predictive utility of behavior-based interviewing compared with traditional interviewing in the selection of radiology residents. *Investigative Radiology*. 27(5), pp.385-389.

Amos, D.E. and Massagli, T.L., 1996. Medical school achievements as predictors of performance in a physical medicine and rehabilitation residency. *Academic Medicine.* 71(6), pp.678-80.

Archer, J. et al., 2008. mini-PAT (Peer Assessment Tool): A valid component of a national assessment program in the UK? *Adv in Health Science Education* 13(2), pp.181-192

Arthur, W., Jr., Day, E.A., McNelly, T.L. & Edens, P.S., 2003. A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology.* 56, pp.125-154.

Baker, J.D 3rd. et al., 1993. Selection of anaesthesiology residents. *Academic Medicine*. 68(2), pp.161-3.

Balentine, J. Gaeta, T. & Spevack, T., 1999. Evaluating applicants to emergency medicine residency programs. *Journal of Emergency Medicine*. 17(1), pp.131-4.

Barclay, J. M., 1999. Employee selection: A question of structure. *Personnel Review.* 28 (1/2), pp.134-151.

Barrick, M.R. and Zimmerman, R.D., 2005. Reducing voluntary, avoidable turnover through selection. *Journal of Applied Psychology*. 90(1), pp.159-66.

Basco, W.T Jr., Gilbert, G.E., Chessman, A.W. & Blue, A.V., 2000. The ability of a medical school admission process to predict clinical performance and patients' satisfaction. *Academic Medicine*. 75(7), pp.743-7.

Basco, W.T Jr. et al., 2008. Medical school application interview score has limited predictive validity for performance on a fourth year clinical practice examination. *Advances in Health Sciences Education*. 13(2), pp.151-62.

Bell, J.G., Kanellitsas, I. & Shaffer, L., 2002. Selection of obstetrics and gynaecology residents on the basis of medical school performance. *American Journal of Obstetrics & Gynaecology*. 186(5), p.1091-4.

Berner, E.S., Brooks, C.M. & Erdmann, J.B., 1993. Use of the USMLE to select residents. *Academic Medicine*. 68(10), pp.753-9.

Bindal, T., Wall, D. & Goodyear, H.M., 2007. Performance of paediatric Senior House Officers following changes in recruitment. *Medical Teacher*. 29(5), pp.501-3.

Black, K.P., Abzug, J.M. & Chinchilli, V.M., 2006. Orthopaedic in-training examination scores: a correlation with USMLE results. *Journal of Bone & Joint Surgery - American*. 88(3), pp.671-6.

Bland, A.C., Kreiter, C.D. & Gordon, J.A., 2005. The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine*. 80(4), pp.395-9.

Borowitz, S.M., Saulsbury, F.T. & Wilson, W.G., 2000. Information collected during the residency match process does not predict clinical performance. *Archives of Pediatric and Adolescent Medicine*. 154, pp.256-260.

Boursicot, K.A., Roberts, T.E. & Pell, G., 2006. Standard setting for clinical competence at graduation from medical school: a comparison of passing scores across five medical schools. *Advances in Health Sciences Education*. 11(2), pp.173-83.

Boyse, T.D. et al., 2002. Does medical school performance predict radiology resident performance? *Academic Radiology*. 9, pp.437-445.

Brothers, T.E. and Wetherholt, S., 2007. Importance of the faculty interview during the resident application process. *Journal of Surgical Education*. 64(6), pp.378-85.

Brown, E., Rosinski, E.F. & Altman, D.F., 1993. Comparing medical school graduates who perform poorly in residency with graduates who perform well. *Academic Medicine*. 68(10), pp.806-8.

Brownell, K., Lockyer, J., Collin, T. & Lemay, J.F., 2007. Introduction of the multiple mini interview into the admissions process at the University of Calgary: acceptability and feasibility. *Medical Teacher.* 29(4), pp.394-6.

Callaghan, C.A. et al., 2000. Validity of faculty ratings of students' clinical competence in core clerkships in relation to scores on licensing examinations and supervisors' ratings in residency'. *Academic Medicine*. 75, pp.S71-73.

Campion, M.A., Palmer, D. K. & Campion, J. E., 1997. A review of structure in the selection interview. *Personnel Psychology*, 50, pp.655-702.

Carrothers, R.M., Gregory, S.W Jr. & Gallagher, T.J., 2000. Measuring emotional intelligence of medical school applicants. *Academic Medicine*. 75(5), pp.456-63.

Case, S.M. and Swanson, D.B., 1993. Validity of NBME Part I and Part II scores for selection of residents in orthopaedic surgery, dermatology and preventative medicine. *Academic Medicine*. 68, pp.S51-S56.

Chamberlain, S.E. and Searle, J., 2005. Assessing suitability for a problem-based learning curriculum: evaluating a new student selection instrument. *Medical Education*. 39(3), pp.250-7.

Chamberlain, T.C., Catano, V.M. & Cunningham, D.P., 2005. Personality as a predictor of professional behavior in dental school: comparisons with dental practitioners. *Journal of Dental Education*. 69(11), pp.1222-37.

Chamorro-Premuzic, T., Furnham, A. & Ackerman, P.L., 2006. Incremental validity of the typical intellectual engagement scale as predictor of different academic performance measures. *Journal of Personality Assessment*. 87(3), pp.261-8.

Chapman, D,S., Uggerslev, K.L. & Webster, J., 2003. Applicant reactions to face-to-face and technology-mediated interviews: a field investigation. *Journal of Applied Psychology*. 88(5), pp.944-53.

Coates, H., 2008. Establishing the criterion validity of the Graduate Medical School Admissions Test (GAMSAT). *Medical Education*. 42(10), pp.999-1006.

Cohen-Schotanus J. et al., 2006. The predictive validity of grade point average scores in a partial lottery medical school admission system. *Medical Education*. 40(10), pp.1012-9.

Collins, J.P., White, G.R., Petrie, K.J. & Willoughby, E.W., 1995. A structured panel interview and group exercise in the selection of medical students. *Medical Education*. 29(5), pp.332-336.

Courneya C. et al., 2005. Medical student selection: choice of a semi-structured panel interview or an unstructured one-on-one interview. *Medical Teacher*. 27(6), pp.499-503.

Crane, J.T. and Ferraro, C.M., 2000. Selection criteria for emergency medicine residency applicants. *Academic Emergency Medicine*. 7(1):54-60.

DaRosa, D.A. and Folse, R., 1991. Evaluation of a system designed to enhance the resident selection process. *Surgery*. 109(6), pp.715-21.

DeLisa, J.A., Jain, S.S. & Campagnolo, D.I., 1994. Factors used by physical medicine and rehabilitation residency training directors to select their residents. *American Journal of Physical Medicine & Rehabilitation*. 73(3), pp.152-6.

Dewhurst, N.G. et al., 2007. Performance in the MRCP(UK) Examination 2003-4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. *BMC Medicine*. 5:8.

Dipboye, R. L., 1994. Structured and unstructured selection interviews: Beyond the job-fit model. In G. R. Ferris (Ed.), *Research in personnel and human resources management.* 12 pp. 79-123. Greenwich, CT: JAI Press.

Dirschl, D.R., 2002. Scoring of orthopaedic residency applicants: is a scoring system reliable? *Clinical Orthopaedics & Related Research*. (399), pp.260-4.

Dirschl, D.R., Campion, E.R. & Gilliam, K., 2006. Resident selection and predictors of performance: can we be evidence based? *Clinical Orthopaedics & Related Research*. 449, pp.44-9.

Dirschl, D.R. et al., 2002. Correlating selection criteria with subsequent performance as residents. *Clinical Orthopaedics & Related Research*. (399), pp.265-71.

Dodson, M. et al., 2009. The multiple mini-interview: how long is long enough? *Medical Education*. 43(2), pp.168-74.

Donnon, T. and Paolucci, E.O., 2008. A generalizability study of the medical judgment vignettes interview to assess students' noncognitive attributes for medical school. *BMC Medical Education*. 8, p.58.

Dore, K.L. et al., 2006. Medical school admissions: enhancing the reliability and validity of an autobiographical screening tool. *Academic Medicine*. 81(10 Suppl), pp.S70-3.

Driessen, E.W. et al., 2006. Validity of portfolio assessment: which qualities determine ratings?. *Medical Education*. 40(9), pp.862-6.

Durning, S.J. et al., 2005. The feasibility, reliability, and validity of a program director's (supervisor's) evaluation form for medical school graduates. *Academic Medicine*. 80(10), pp.964-8.

Elam, C.L. et al., 2002. Review, deliberation, and voting: a study of selection decisions in a medical school admission committee. *Teaching & Learning in Medicine*. 14(2), pp.98-103.

Elam, C.L., Stratton, T.D., Wilson, J.F. & Scott, K.L. 2002. How admission committees decide: influence of committee members' experience and applicants' academic characteristics. *Academic Medicine*. 77(10 Suppl), pp.S26-8.

Elliott, S.L. and Epstein, J., 2005. Selecting the future doctors: the role of graduate medical programmes. *Internal Medicine Journal*. 35(3), pp.174-7.

Eva, K.W., Rosenfeld, J., Reiter, H.I. & Norman, G.R., 2004 (a). An admissions OSCE: the multiple mini-interview. *Medical Education*. 38(3), pp.314-26.

Eva, K.W., Reiter, H.I., Rosenfeld, J. & Norman, G.R., 2004 (b). The ability of the multiple mini-interview to predict preclerkship performance in medical school. *Academic Medicine*. 79(10 Suppl), pp.S40-2.

Eva, K.W., Reiter, H.I., Rosenfeld, J. & Norman, G.R., 2004 (c). The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. *Academic Medicine*. 79(6), pp.602-9.

Evans, P. and Wen, F.K., 2007. Does the medical college admission test predict global academic performance in osteopathic medical school? *Journal of the American Osteopathic Association*. 107(4), pp.157-62.

Fields, H.W., Fields, A.M. & Beck, F.M., 2003. The impact of gender on high-stakes dental evaluations. *Journal of Dental Education*. 67(6), pp.654-60.

Frantsve, L.M., Laskin, D.M. & Auerbach, S.M., 2003. Personality and gender influences on faculty ratings and rankings of oral and maxillofacial surgery residency applicants. *Journal of Dental Education*. 67(11), pp.1252-9.

Furnham, Adrian., 2008. HR professionals' beliefs about, and knowledge of, assessment techniques and psychometric tests. *International Journal of Selection and Assessment*. 16(3), pp.300-305.

Galazka, S.S., Kikano, G.E. & Zyzanski, S., 1994. Methods of recruiting and selecting residents for U.S. family practice residencies *Academic Medicine*. 69(4), pp.304-6.

Garman, A. N. and Lesowitz, T., 2005. Research update: Interviewing candidates for leadership roles. *Consulting Psychology Journal: Practice and research.* 57(4), pp.266-273.

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C., III. & Bentson, C., 1987. Meta-analysis of assessment center validity. *Journal of Applied Psychology Monograph, 72*, pp.493-511.

George, J.M., Young, D. & Metz, E.N., 1989. Evaluating selected internship candidates and their subsequent performances. *Academic Medicine*. 64(8), pp.480-2.

Gilbart, M.K., Cusimano, M.D. & Regehr, G., 2001. Evaluating surgical resident selection procedures. *American Journal of Surgery*. 181(3), pp.221-5.

Glick, S.M., 2000. Selection for entry to medicine and specialist training. *Medical Teacher*. 22(5), pp.443-447.

Goho, J. and Blackman, A., 2006. The effectiveness of academic admission interviews: an exploratory meta-analysis. *Medical Teacher*. 28(4), pp.335-40.

Goodyear, H.M., Jyothish, D., Diwakar, V. & Wall, D., 2007. Reliability of a regional junior doctor recruitment process. *Medical Teacher*. 29(5), pp.504-6.

Gorman, D., Monigatti J. & Poole, P., 2008. On the case for an interview in medical student selection. *Internal Medicine Journal*. 38(8), pp.621-623.

GMC. 2003. Tomorrow's Doctor. London: General Medical Council

GMC. 2007. The New Doctor 2007. London: General Medical Council

Griffin, B., Harding, D.W., Wilson, I.G. & Yeomans, N.D., 2008. Does practice make perfect? The effect of coaching and retesting on selection tests used for admission to an Australian medical school.[erratum appears in Med J Aust. 2008 Oct 6;189(7):416]. *Medical Journal of Australia*. 189(5), pp.270-3.

Groves, M.A., Gordon, J. & Ryan, G., 2007. Entry tests for graduate medical programs: is it time to re-think? *Medical Journal of Australia*. 186(3), pp.120-3.

Guion, R. A., 1998. Assessment, Measurement, and Prediction for Personnel Decisions. Mahwah, NJ: Lawrence Erlbaum Associates.

Gunderman, R.B. and Jackson, V.P., 2000. Are NBME examination scores useful in selecting radiology residency candidates? *Academic Radiology*. 7(8), pp.603-6.

Hall, F.R., Regan-Smith, M. & Tivnan, T., 1992. Relationship of medical students' admission interview scores to their dean's letter ratings. *Academic Medicine*. 67(12), pp.842-5.

Hall, F.R. and Bailey, B.A., 1992. Correlating students' undergraduate science GPAs, their MCAT scores, and the academic caliber of their undergraduate colleges with their first-year academic performances across five classes at Dartmouth Medical School. *Academic Medicine*. 67(2), pp.121-3.

Halley, M.C., Lalumandier, J.A., Walker, J.D. & Houston, J.H., 2008. A regional survey of dentists' preferences for hiring a dental associate. *Journal of the American Dental Association*. 139(7), pp.973-9.

Hamdy, H. et al., 2006. BEME systematic review: predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher*. 28(2), pp.103-16.

Hanson, M.D., Dore, K.L., Reiter, H.I. & Eva, K.W., 2007. Medical school admissions: revisiting the veracity and independence of completion of an autobiographical screening tool. *Academic Medicine*. 82(10 Suppl), pp.S8-S11.

Harasym, P.H., Woloschuk, W., Mandin, H. & Brundin-Mather, R., 1996. Reliability and validity of interviewers' judgments of medical school candidates. *Academic Medicine*. 71(1 Suppl), pp.S40-2.

Harris, S. and Owen, C., 2007. Discerning quality: using the multiple mini-interview in student selection for the Australian National University Medical School. *Medical Education*. 41(3), pp.234-41.

Hayden, S.R., Hayden, M. & Gamst, A., 2005. What characteristics of applicants to emergency medicine residency programs predict future success as an emergency medicine resident? *Academic Emergency Medicine*. 12(3), pp,206-10.

Heintze, U., Radeborg, K., Bengtsson, H. & Stenlaas, A., 2004. Assessment and evaluation of individual prerequisites for dental education. *European Journal of Dental Education*. 8(4), pp.152-60.

Hemaida, R.S. and Kalb, E., 2001. Using the analytic hierarchy process for the selection of first-year family practice residents. *Hospital Topics*. 79(1), pp.11-5.

Hoad-Reddick, G. and Macfarlane, T.V., 1999. Organising the introduction of, and evaluating interviewing in, an admissions system. *European Journal of Dental Education*. 3(4), pp.172-9.

Hojat, M. et al., 2000. A validity study of the writing sample section of the medical college admission test. *Academic Medicine*. 75(10 Suppl), pp.S25-7.

Hojat, M. et al., 1993. Students' psychosocial characteristics as predictors of academic performance in medical school *Academic Medicine*. 68(8), pp.635-7.

Hojat, M., Gonnella, J.S., Veloski, J.J. & Erdmann, J.B., 1993. Is the glass half full or half empty? A re-examination of the association between assessment measures during medical school and clinical competence after graduation. *Academic Medicine*. 68, pp.S69-S76.

Hough, L.M. and Oswald, F.L., 2000. Personnel selection: looking toward the future--remembering the past. *Annual Review of Psychology*. 51, pp.631-64.

Huffcutt, A.I. and Woehr, D.J., 1999. Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior*. 20(4), pp.549-560.

Huffcutt, A. I. and Arthur, W., Jr. 1994. Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology,* 79, pp.184-190.

Huffcutt, A. I., Conway, J.M., Roth, P.L. & Klehe, U.C., 2004. The impact of job complexity and study design on situational and behaviour description interview validity. *International Journal of Selection and Assessment.* 12(3), pp.262-273.

Humphrey, S. et al., 2008. Multiple mini-interviews: opinions of candidates and interviewers. *Medical Education*. 42(2), pp.207-13.

Hysong, S.J. and Dipboye, R.L., 1999. Individual differences in applicant's reactions to employment interview elements. Paper presented at the 14th Annual Conference of the Society for Industrial/Organizational Psychology. Atlanta, Georgia. As cited in Moscoso S. (2000). Selection interview: A review of validity evidence, adverse impact and applicant reactions.

Illing J, Morrow G, Kergon C, Burford B, Spencer J, Peile E, et al. (2008) How prepared are medical graduates to begin practice? A comparison of three diverse UK medical schools. Final summary and conclusions for the GMC Education Committee, http://www.gmc-uk.org/about/research/REPORT%20-preparedness%20of%20medical%20grads.pdf [accessed 14.4.09]

Universities UK. (2007) Beyond the honours degree classification: The Burgess Group final report [http://www.universitiesuk.ac.uk/Publications/Documents/Burgess_final.pdf - accessed 17 April 2009]

International Task Force on Assessment Center Guidelines 2000. *Guidelines and ethical considerations for assessment center operations*. Endorsed by the 28th International Congress on Assessment Center Methods, May 2000, San Francisco, CA.

Iramaneerat, C., 2006. Predicting academic achievement in the medical school with high school grades. *Journal of the Medical Association of Thailand*. 89(9), pp.1497-505.

Janis, J.E. and Hatef, D.A., 2008. Resident selection protocols in plastic surgery: a national survey of plastic surgery program directors. *Plastic & Reconstructive Surgery*. 122(6), pp.1929-39; discussion pp.1940-1.

Johnson, E.K. and Edwards, J.B., 1991. Current practices in admissions interviews as U.S. medical schools. *Academic Medicine.* 7, pp.408-412.

Joyner, P.U., Cox, W.C., White-Harris, C. & Blalock, S.J., 2007. The structured interview and interviewer training in the admissions process. *American Journal of Pharmaceutical Education*. 71(5), pp.83.

Khan, M.J. et al., 2001. Residency program director evaluations do not correlate with performance on a required 4th-year objective structured clinical education. *Teaching & Learning in Medicine*. 13, pp.9-12.

Kingsley K. Sewell J. Ditmyer M. O'Malley S. Galbraith GM. Creating an evidence-based admissions formula for a new dental school: University of Nevada, Las Vegas, School of Dental Medicine. Journal of Dental Education. 71(4):492-500, 2007 Apr.

Kinicki, A.J., Lockwood, C.A., Hom, P.W. & Griffeth, R.W., 1990. Interviewer predictions of applicant qualifications and interviewer validity: aggregate and individual analyses. *Journal of Applied Psychology*. 75(5), pp.477-86.

Knights, J.A. and Kennedy, B.J., 2006. Medical school selection: Screening for dysfunctional tendencies. *Medical Education*. 40(11), pp.1058-64.

Kogan, J.R. and Shea, J.A., 2005. Psychometric characteristics of a write-up assessment form in a medicine core clerkship. *Teaching & Learning in Medicine*. 17(2), pp.101-6.

Kreiter, C. and Solow, C., 2002. A statistical technique for the development of an alternate list when using constrained optimization to make admission decisions. *Teaching & Learning in Medicine*. 14(1), pp.29-33.

Kreiter, C.D., Yin, P., Solow, C. & Brennan, R.L., 2004. Investigating the reliability of the medical school admissions interview. *Advances in Health Sciences Education*. 9(2), pp.147-59.

Kreiter, C.D., Stansfield, B., James, P.A. & Solow, C., 2003. A model for diversity in admissions: a review of issues and methods and an experimental approach. *Teaching & Learning in Medicine*. 15(2), pp.116-22.

Kulatunga-Moruzi, C. and Norman, G.R., 2002 (a). Validity of admissions measures in predicting performance outcomes: the contribution of cognitive and non-cognitive dimensions. *Teaching & Learning in Medicine*. 14(1), pp.34-42.

Kulatunga-Moruzi, C. and Norman, G.R., 2002 (b). Validity of admissions measures in predicting performance outcomes: a comparison of those who were and were not accepted at McMaster. *Teaching & Learning in Medicine*. 14(1), pp.43-8.

Lee, A.G. et al., 2008. Re-engineering the Resident Applicant Selection Process in Ophthalmology: A Literature Review and Recommendations for Improvement. *Survey of Ophthalmology*. 53(2), pp.164-176.

Lemay, J.F., Lockyer, J.M., Collin, V.T. & Brownell, A.K., 2007. Assessment of non-cognitive traits through the admissions multiple mini-interview. *Medical Education*. 41(6), pp.573-9.

Lievens, F., Buyse, T. & Sackett, P.R., 2005. The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*. 90(3), pp.442-52.

Loftus, L.S., Arnold, L., Willoughby, T.L & Connolly, A., 1992. First-year residents' performance compared with their medical school class ranks as determined by three ranking systems. *Academic Medicine*. 67, pp.319-323.

Lumb, A.B. and Vail, A., 2004. Comparison of academic, application form and social factors in predicting early performance on the medical course. *Medical Education*. 38(9), pp.1002-5.

Lurie, S.J., Lambert, D.R. & Grady-Weliky, T.A., 2007. Relationship between dean's letter rankings and later evaluations by residency program directors. *Teaching & Learning in Medicine*. 19(3), pp.251-6.

Market, R.J., 1993. The relationship of academic measures in medical school to performance after graduation. *Academic Medicine*. 68, pp.S31-S34.

Marley, J. and Carman, I., 1999. Selecting medical students: A case report of the need for change. *Medical Education*. 33(6), pp.455-459.

Marrin, M.L., McIntosh, K.A., Keane, D. & Schmuck, M.L., 2004. Use of the paired-comparison technique to determine the most valued qualities of the McMaster Medical Programme Admissions Process. *Advances in Health Sciences Education*. 9(2), pp.129-35.

McCarey, M., Barr, T. & Rattray, J., 2007. Predictors of academic performance in a cohort of pre-registration nursing students. *Nurse Education Today*. 27(4), pp.357-64.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L. & Maurer, S. D., 1994. The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology,* 79, pp.599-616.

McManus, I.C. et al., 2005. Intellectual aptitude tests and A levels for selecting UK school leaver entrants for medical school. *British Medical Journal*. 331(7516), pp.555-559.

McManus, I,C. et al., 2005. Unhappiness and dissatisfaction in doctors cannot be predicted by selectors from medical school application forms: a prospective, longitudinal study. *BMC Medical Education.* 5, p.38.

McManus, I.C. et al., 2003. A levels and intelligence as predictors of medical careers in UK doctors: 20 year prospective study. *British Medical Journal*. 327(7407), pp.139-42.

McManus, I.C., 1998. Factors affecting likelihood of applicants being offered a place in medical schools in the United Kingdom in 1996 and 1997: prospective study. *British Medical Journal.* 317, pp.1111-1117

Metro, D.G., Talarico, J.F., Patel, R.M. & Wetmore, A.L., 2005. The resident application process and its correlation to future performance as a resident. *Anesthesia & Analgesia*. 100(2), pp.502-5.

Miles, W.S., Shaw, V. & Risucci, D., 2001. The role of blinded interviews in the assessment of surgical residency candidates. *American Journal of Surgery*. 182(2), pp.143-6.

Miller, G.E., 1990. The assessment of clinical skills/competence/performance. *Academic Medicine.* 65, pp.S63-67.

Milne, C.K. et al., 2003. Residents as members of intern selection committees: can they partially replace faculty? *Teaching & Learning in Medicine*. 15(4), pp.242-6.

Milne, C.K., Bellini, L.M. & Shea, J.A., 2001. Applicants' perceptions of the formal faculty interview during residency recruitment. *Academic Medicine*. 76(5), p.501.

Murden, R.A., Way, D.P., Hudson, A. & Westman, J.A., 2004. Professionalism deficiencies in a first-quarter doctor-patient relationship course predict poor clinical performance in medical school. *Academic Medicine*. 79(10 Suppl), pp.S46-8.

National Patient Safety Agency. *National Clinical Assessment Service: Analysis of the First Four Years' Referral Data*, July 2006

Nayer, M., 1992. Admission criteria for entrance to physiotherapy schools: how to choose among many applicants. *Physiotherapy Canada*, 44, pp.41-46.

Nowacek, G.A., Bailey. B,A. & Sturgill, B.C., 1996. Influence of the interview on the evaluation of applicants to medical school. *Academic Medicine*. 71(10), pp.1093-5.

Olawaiye, A., Yeh, J. & Withiam-Leitch, M., 2006. Resident selection process and prediction of clinical performance in an obstetrics and gynecology program. *Teaching & Learning in Medicine*. 18(4), pp.310-5.

O'Neill, L.D. et al., 2009. Generalisability of a composite student selection programme. *Medical Education*. 43(1), pp.58-65.

Oosterveld, P. and ten Cate, O., 2004. Generalizability of a study sample assessment procedure for entrance selection for medical school. *Medical Teacher*. 26(7), pp.635-9.

Otero, H.J., Erturk, S.M., Ondategui-Parra, S. & Ros, P.R., 2006. Key criteria for selection of radiology residents: results of a national survey. *Academic Radiology*. 13(9), pp.1155-64.

Paolo, A.M. and Bonaminio, G.A., 2003. Measuring outcomes of undergraduate medical education: residency directors' ratings of first-year residents. *Academic Medicine*. 78, pp.90-95.

Paolo, A.M. et al., 2006. A comparison of students from main and alternate admission lists at one school: the potential impact on student performance of increasing enrollment. *Academic Medicine*. 81(9), pp.837-41.

Papp, K.K., Polk, H.C Jr. & Richardson, J.D., 1997. The relationship between criteria used to select residents and performance during residency. *American Journal of Surgery*. 173(4), pp.326-9.

Park, S.E., Susarla, S.M. & Massey, W., 2006. Do admissions data and NBDE Part I scores predict clinical performance among dental students? *Journal of Dental Education*. 70(5), pp.518-24.

Patrick, L.E., Altmaier, E.M., Kuperman, S. & Ugolini, K., 2001. A structured interview for medical school admission, Phase 1: initial procedures and results. *Academic Medicine*. 76(1), pp.66-71.

Patterson, F. et al., 2000. A competency model for general practice: implications for selection, training, and development. *British Journal of General Practice*. 50(452), pp.188-93.

Patterson, F., Ferguson, E., Norfolk, T.& Lane, P., 2005. A new selection system to recruit general practice registrars: preliminary findings from a validation study. *British Medical Journal*. 330(7493), pp.711-4.

Patterson, F., Ferguson, E. & Thomas, S., 2008. Using job analysis to identify core and specific competencies: Implications for selection and recruitment. *Medical Education*. 42(12), pp.1195-1204.

Patterson, F. and Ferguson, E., 2007. Selection for medical education and training. ASME monograph, Understanding medical education.

Pearson, S.A., Rolfe, I.E. & Henry, R.L., 1998. The relationship between assessment measures at Newcastle Medical School (Australia) and performance ratings during internship. *Medical Education*. 32, pp.40-45.

Peng, R., Khaw, H.H. & Edariah, A.B., 1995. Personality and performance of preclinical medical students. *Medical Education*. 29(4), pp.283-8.

Peskun, C., Detsky, A. & Shandling, M., 2007. Effectiveness of medical school admissions criteria in predicting residency ranking four years later. *Medical Education*. 41(1), pp.57-64.

Pilon, S. and Tandberg, D., 1997. Neural network and linear regression models in residency selection. *American Journal of Emergency Medicine*. 15(4), pp.361-4.

Poole, A., Catano, V.M. & Cunningham, D.P., 2007. Predicting performance in Canadian dental schools: the new CDA structured interview, a new personality assessment, and the DAT. *Journal of Dental Education*. 71(5), pp.664-76.

Posthuma RA, Morgeson FP, Campion, MA. 2002. Beyond employment interview validity: A comprehensive narrative reviewof recent research and trends over time. *Personnel Psychology* 55; 1-81

Price, S.S. et al., 2008. Increasing minority enrolment utilizing dental admissions workshop strategies. *Journal of Dental Education*. 72(11), pp.1268-76.

Rabinowitz, H.K. & Hojat, M.A., 1989. Comparison of the modified essay question and multiple choice question formats: their relationship to clinical performance. Family Medicine. 21, pp.364-367.

Randall, R., Davies, H., Patterson, F. & Farrell, K., 2006. Selecting doctors for postgraduate training in paediatrics using a competency based assessment centre. *Archives of Disease in Childhood*. 91(5), pp.444-8.

Randall, R., Stewart, P., Farrell, K. & Patterson, F., 2006. Using an assessment centre to select doctors for postgraduate training in obstetrics and gynaecology. *The Obstetrician & Gynaecologist*. 8, pp257-262.

Rees, C. and Sheard, C., 2004. UG medical students' views about a reflective portfolio assessment of their communication skills learning. *Medical Education*. 38(2), pp.125-128.

Reiter, H.I., Eva, K.W., Rosenfeld, J. & Norman, G.R., 2007. Multiple mini-interviews predict clerkship and licensing examination performance. *Medical Education*. 41(4), pp.378-84.

Richards, J.M., Taylor, C.W. & Price, P.B., 1962. The prediction of medical intern performances. *Journal of Applied Psychology*. 46, pp.142-146.

Rifkin, W.D. and Rifkin, A., 2005. Correlation between housestaff performance on the United States Medical Licensing Examination and standardized patient encounters. *Mount Sinai Journal of Medicine*. 72(1), pp.47-9.

Roberts, C. et al., 2008. Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Medical Education*. 42(4), pp.396-404.

Rosenfeld, J.M., Reiter, H.I., Trinh, K. & Eva, K.W., 2008. A cost efficiency comparison between the multiple mini-interview and traditional admissions interviews. *Advances in Health Sciences Education*. 13(1), pp.43-58.

Rutala, P.J. et al., 1992. Validity studies using standardized-patient examinations: standardized patient potpourri. *Academic Medicine*. 67, pp.S60-S62.

Sackett, P.R. and Lievens, F., 2008. Personnel selection. *Annual Review of Psychology*. 59, pp.419-50.

Sadler, J., 2003. Effectiveness of student admission essays in identifying attrition. *Nurse Education Today*. 23(8), pp.620-7.

Salvatori, P., 2001. Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education*. 6(2), pp.159-75.

Savoldelli, G.L., 2006. Evaluation of patient simulator performance as an adjunct to the oral examination for senior anaesthesia residents. *Anaesthesiology*. 104(3), pp.475-81.

Shaw, D.L., Martz, D.M., Lancaster, C.J. & Sade, R.M., 1995. Influence of medical school applicants' demographic and cognitive characteristics on interviewers' ratings of noncognitive traits. *Academic Medicine*. 70(6), pp,532-6.

Silver, B. and Hodgson, C.S., 1997. Evaluating GPAs and MCAT scores as predictors of NBME I and clerkship performances based on students' data from one undergraduate institution.[see comment]. *Academic Medicine*. 72(5), pp.394-6, May.

Smilen, S.W., Funai, E.F. & Bianco, A.T., 2001. Residency selection: should interviewers be given applicants' board scores? *American Journal of Obstetrics & Gynecology*. 184(3), pp.508-13.

Smith, D.B., Hanges, P.J. & Dickson, M.W., 2001. Personnel selection and the five-factor model: reexamining the effects of applicant's frame of reference. *Journal of Applied Psychology*. 86(2), pp.304-15.

Smith, S.R., 1991. Medical school and residency performances of students admitted with and without an admission interview. *Academic Medicine*. 66(8), pp.474-6.

Smith, S.R., 1993. Correlations between graduates’ performances as first-year residents and their performances as medical students. Academic Medicine. 68, pp.633-634.

Sosenko, J., Stekel, K.W., Soto, R. & Gelbard, M., 1993. NBME Examination Part I as a predictor of clinical and ABIM certifying examination performances. *Journal of General Internal Medicine*. 8, pp.86-88.

Spafford, M.M. and Beal, P.I., 1999. Interview expectations and experiences of women and men applying to an optometry program. *Optometry & Vision Science*. 76(7), pp.500-10.

Spina, A.M., Smith, T.A., Marciani, R.D. & Marshall, E.O., 2000. A survey of resident selection procedures in oral and maxillofacial surgery. *Journal of Oral & Maxillofacial Surgery*. 58(6), pp.660-6; discussion pp.666-7.

Stacey, D.G. and Whittaker, J.M., 2005. Predicting academic performance and clinical competency for international dental students: seeking the most efficient and effective measures. *Journal of Dental Education*. 69(2), pp.270-80.

Stansfield, R.B. and Kreiter, C.D., 2007. Conditional reliability of admissions interview ratings: extreme ratings are the most informative. *Medical Education*. 41(1), pp.32-8.

Stratton, T.D., Elam, C.L., Murphy-Spencer, A.E. & Quinlivan, S.L., 2005. Emotional intelligence and clinical skills: preliminary results from a comprehensive clinical performance examination. *Academic Medicine*. 80(10 Suppl), pp.S34-7.

Suhayda, R., Hicks F. & Fogg, L., 2008. A decision algorithm for admitting students to advanced practice programs in nursing. *Journal of Professional Nursing*. 24(5), pp.281-4.

Swide, C., Lasater, K. & Dillman, D., 2009. Perceived predictive value of the Medical Student performance Evaluation (MSPE) in anaesthesiology resident selection *Journal of Clinical Anaesthesia*. 21(1), pp.38-43.

Tamblyn, R. et al., 2007. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *The Journal of the American Medical Association*. 298(9), pp.993-1001.

Taylor, M.L. et al., 2005. The relationship between the National Board of Medical Examiners' prototype of the Step 2 clinical skills exam and interns' performance. *Academic Medicine*. 80(5), pp.496-501.

ten Cate, O. and Smal, K., 2002. Educational assessment center techniques for entrance selection in medical school. *Academic Medicine*. 77(7), pp.737.

The Europe Unit. (2006) Guide to The Diploma Supplement. [http://www.europeunit.ac.uk/sites/europe_unit2/resources/Guide%20to%20the%20Diploma%20Supplement.pdf - accessed 17 April 2009]

Thordarson, D.B. et al., 2007. Resident selection: how we are doing and why? *Clinical Orthopaedics & Related Research*. 459, pp.255-9.

Tran, T. and Blackman, M.C., 2006. The dynamics and validity of the group selection interview. *Journal of Social Psychology*. 146(2), pp.183-201.

Travis, C., Taylor, C.A. & Mayhew, H.E., 1999. Evaluating residency applicants: stable values in a changing market. *Family Medicine*. 31(4), pp.252-6.

Trewby, P.N., 2005. Assisting international medical graduates applying for their first post in the UK: what should be done? *Clinical Medicine*. 5(2), pp.126-32.

Turner, N.S. et al., 2006. A quantitative composite scoring tool for orthopaedic residency screening and selection. *Clinical Orthopaedics and Related Research*. (449), pp.50-55.

Umansky, J., Taub, P., Lorenz, H.P. & Kawamoto, H.K., 2003. Factors determining the ultimate fate of a plastic surgery applicant. *Plastic & Reconstructive Surgery.* 111(3), pp.981-4. discussion pp.985-6.

Utzman, R.R., Riddle, D,L. & Jewell, D,V., 2007. Use of demographic and quantitative admissions data to predict performance on the national physical therapy examination. *Physical Therapy*. 87(9), pp.1181-93.

Vu, N.V. et al., 1992. Six years of comprehensive, clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine. *Academic Medicine*. 67, p.42-50.

Westwood, M.A., 2008. Applicants regard structured interviews as a fair method of selection: an audit of candidates. *Journal of the Royal Society of Medicine*. 101(5), pp.252-8.

Wiesner, W. H. and Cronshaw, S. F., 1988. A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology,* 61, pp.275-290.

Wilkinson, T.J. and Frampton, C.M., 2004 Comprehensive undergraduate medical assessments improve prediction of clinical performance. *Medical Education*. 38, pp.1111-1116.

Wood, P.S. et al., 1990. A prospective study of cognitive and noncognitive selection criteria as predictors of resident performance. *Investigative Radiology*. 25(7), pp.855-9.

Yindra, K.K., Rosenfield, P.S. & Donnelly, M.B., 1998. Medical school achievements as predictors of residency performance. *Journal Medical Education*. 63, pp.356-363.

Ziv, A. et al., 2008. MOR: a simulation-based assessment centre for evaluating the personal and interpersonal qualities of medical school candidates. *Medical Education*. 42(10), pp.991-8.

## 15. Appendix A

### Search Strategy – full Medline search

Database: Ovid MEDLINE(R) <1950 to February Week 4 2009>
Search Strategy:
--------------------------------------------------------------------------------
1 exp Education, Medical, Undergraduate/ (15925)
2 "Internship and Residency"/ (26285)
3 Education, Medical, Graduate/ (16436)
4 Foreign Medical Graduates/ (2606)
5 Students, Medical/ (14881)
6 Schools, Medical/ (17508)
7 School Admission Criteria/ (3321)
8 College Admission Test/ (316)
9 Licensure, Medical/ (3215)
10 applicant$.ab,ti. (3630)
11 (admission$ adj (criteri$ or standard$ or measure$ or factor$)).mp. [mp=title, original title, abstract, name of substance word, subject heading word] (3941)
12 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 (84394)
13 career choice/ (12610)
14 Educational Status/ (27305)
15 exp "Predictive Value of Tests"/ (88892)
16 predict$.ab,ti. (538733)
17 (13 or 14) and (15 or 16) (3783)
18 17 or 12 (87854)
19 Interview/ (19557)
20 Interview, Psychological/ (8830)
21 interview$.ab,ti. (138584)
22 21 or 19 or 20 (148205)
23 (rank$ adj2 exam$).mp. [mp=title, original title, abstract, name of substance word, subject heading word] (170)
24 (national adj2 "clinical exam$").mp. [mp=title, original title, abstract, name of substance word, subject heading word] (7)
25 24 or 23 (177)
26 "free text".mp. [mp=title, original title, abstract, name of substance word, subject heading word] (714)
27 freetext.mp. [mp=title, original title, abstract, name of substance word, subject heading word] (12)
28 "white space".mp. [mp=title, original title, abstract, name of substance word, subject heading word] (34)
29 personal statement$.mp. [mp=title, original title, abstract, name of substance word, subject heading word] (50)
30 (reflect$ adj statement$).mp. [mp=title, original title, abstract, name of substance word, subject heading word] (3)
31 (personal adj2 reflection$).mp. [mp=title, original title, abstract, name of substance word, subject heading word] (425)
32 27 or 28 or 30 or 26 or 31 or 29 (1238)
33 (record adj2 achievement).mp. [mp=title, original title, abstract, name of substance word, subject heading word] (19)

34 ((portfolio$ or structured) adj2 (report or record)).mp. [mp=title, original title, abstract, name of substance word, subject heading word] (191)
35 33 or 34 (210)
36 Educational Measurement/ (20573)
37 Psychometrics/ (35990)
38 Personnel Selection/ (8863)
39 (assessment adj cent$).mp. [mp=title, original title, abstract, name of substance word, subject heading word] (208)
40 "work sample".mp. [mp=title, original title, abstract, name of substance word, subject heading word] (36)
41 (select$ adj3 (psychometric or psychological or personality or cognitive or composite)).mp. [mp=title, original title, abstract, name of substance word, subject heading word] (1173)
42 38 or 39 or 40 or 36 or 37 or 41 (66188)
43 42 or 35 or 25 or 22 or 32 (210413)
44 18 and 43 (12297)
45 limit 44 to yr="1989 - 2009" (8468)
46 from 45 keep 1-1486 (1486)

***************************