Selection for the Foundation Programme:
A review of available methods

Prepared for the Medical Schools Council

John C McLachlan and Linda L Turnbull

University of Durham
Queen's Campus
Holliday Building
Stockton-on-Tees TS17 6BH
U.K.

Index

Since the Tables are presented in a different format they are to be found in separate files. An appendix is provided which summarises the meaning of key terms such as validity, reliability, utility etc, and provides some general notes on assessment.

Outline Conclusions and Recommendations

The review is divided into decimal numbered sections. Summary Conclusions in each section are numbered according to the section and section level in which they are found: thus *2.3.1* is the first Summary Conclusion in Section 2.3. This allows Conclusions to be matched to Sections. Since this is not a meta-analysis (see Methods), an qualitative description of effect sizes is given, using the conventional Cohen (1988) delimiters: effect sizes below 0.3 are described as *'low'* or *'small'*, around 0.5, are described as *'medium'* or *'moderate'*, and around 0.8 are described as *'high'*, *'good'* or *'strong'*. Since there are almost certainly several constructs under investigation in selection, it is unlikely that any single measure will reach such high reliability that it can be used as the sole selection method. The recommendations are explained in more detail in Section 8.

Summary Conclusions from Sections

2.2.1 There is clear evidence from a variety of sources that performance on national licensing exams is a moderate predictor of performance in later clinical practice, by a variety of measures and outcomes.

2.3.1 The assumption that all graduates from UK medical schools can be ranked nationally on the basis of their local results is not robustly supported by evidence

2.4.1 Strongly held views may militate against a national examination, even if there are arguments in its favour, and its Acceptability in the UK in the current climate is likely to be low.

2.5.1 Test equating avoids some of the problems of 'curriculum paralysis' which may arise from perceptions of the impact of a national testing system.

3.2.1 A considerable body of evidence indicates that traditional interviews have low reliability and validity

3.2.2 A considerable body of evidence suggests that structured interviews have moderate reliability and some evidence of moderate predictive validity.

3.5.3 There is strong and developing evidence that MMI approaches have stronger reliability and predictive validity than other interviewing techniques.

4.1.11 Behaviour which causes concern at undergraduate level has predictive validity for later clinical practice.

4.1.2 Conscientiousness is a significant component of such concerns, and may be possible to measure as an objective scalar property.

4.1.3 Critical Incident forms capture valuable information about adverse events which are below the threshold of referrals to Fitness to Practice Panels.

4.1.4 A narrative summary of information relating to behaviours in Undergraduate programmes would be of value for selection decisions for Foundation places.

4.1.5 Personal Qualities Assessment instruments currently lack the evidence in medical selection which would permit their use for high stakes selection.

5.1 Despite their wide spread use, the predictive validity of portfolios is uncertain

5.2 Under defined conditions of design and training, portfolio assessment may be moderately reliable.

5.3 The cost of administering a portfolio based assessment system comparable across medical schools is likely to be high.

5.4 The general acceptability of portfolio based summative assessment to candidates appears to be low

6.1 Personal statements made in response to structured questions have moderate predictive validity for subsequent performance, but are expensive to score.

7.1 Where different selection methods are combined, scores should be converted to a standardised distribution and weighted according to their reliability.

General Conclusions

1. A national examination for UK medical students, even for ranking purposes, is unlikely to command sufficient support to overcome the logistical difficulties in developing it.
2. A system of Test Equating for cognitive knowledge assessments is practicable, and unlikely to meet with the resistance a national examination would encounter.
3. A system of Test Equating for skills assessments using instruments such as OSCEs poses significant challenges. An assessment centre approach may be more practicable.
4. A consistent method of recording and reporting behaviours relating to professionalism (short of Fitness to Practice proceedings), suitable for use across the U.K, should be developed.
5. Structured interviews, developed after analysis of the job requirements, should be employed in selection in place of the ‘white space’ questions currently employed. These interviews could also take place within assessment centres, and could take the form of Multiple Mini Interviews.

Recommendations

An evidence based, acceptable highly defensible, selection approach for Foundation places could therefore take the following form.

1. Test equating of cognitive knowledge (CK) assessments in each medical school would give candidates a national ‘Knowledge’ rank equivalent.

2. Skills assessment, either through test centres administering a common OSCE programme or, more problematically, through test equating of OSCE skills test within existing curricula, would give each candidate a ‘Skills’ score.

3. Each Medical School Dean would present a structured national report form of ‘behaviours’ for each candidate for review by the Foundation School, based on consistently collected information.

4. Each candidate would receive a structured interview (either as a single interview or through Multiple Mini Interviews) designed to explore those qualities relevant to the job which are amenable to direct observation (e.g. such as interpersonal skills, verbal communication skills, and problem solving). This would generate a national ‘Interview’ score.

5. Foundation Schools could weight these scores according to defined, justifiable and transparent criteria in making selection decisions.

1. Search Methods

The review focused on three databases: Pubmed, PsycINFO and ERIC, which provide access to abstracts from educational, professional, academic, biomedical and life science journals. The search strategy was carried out in four phases related to the four topics identified as potential mechanisms for selection into foundation programmes:

- Interviews
- National ranking examination
- Structured record of achievement (portfolio)
- Application forms based on white space questions

Each phase of the search used 10 core search terms to focus the search process:

- Selection
- Recruit*
- Foundation
- Medical
- Doctor
- Registrar
- Residency
- GP
- General practitioner
- Healthcare

These terms were used in combinations with key words from the identified topics and further refining terms were generated from the search process (identified in brackets):

- Interview*
- Portfolio (PPD, Personal and Professional Development)
- Record of Achievement
- Exam*(national)
- Rank*
- Application form

- White space

Search terms were used in combination to produce manageable lists of article details which were reviewed using the following inclusion and exclusion criteria:

Inclusion Criteria	Exclusion Criteria
Relevant to identified topic	Irrelevant to the identified topic
Relevant to any aspect of utility: validity, reliability, acceptability, educational impact, cost or defensibility.	No aspect of utility addressed
Adequate methodology.	Inadequate methodology identifiable from the abstract.
Informed opinion which adds to the picture of the topic.	Language other than English

Collections of abstracts were merged to remove duplicates and then reviewed in detail in relation to aspects of utility. Relevant citations were followed and added to collections where appropriate. Summary tables were constructed and reference details documented. A total of 129 articles were produced by this process.

Hand searching was carried out from the principal author's Endnote file system, and through key journals in the field (*Medical Education*, *Academic Medicine*, *Medical Teacher* and *Advances in Health Sciences Education*). For key references, citations were searched backwards in time through the HTML files of the document, and forward through the Science Citation Index. Most articles were available electronically, and only in a few cases were inter-library loans required.

In most papers, insufficient information was available to develop a fully consistent information set. Of the desirable characteristics of validity, reliability, educational impact, acceptability and cost, it was rare to find a paper which addressed more than two at any one time. Moreover, even when the same construct was explored, a different approach might be taken to measuring it by different researchers. For instance, 'reliability' might be explored through inter-rater reliability, inter-test reliability, Cronbach's alpha, or Generalisability Theory's G. As a result, although the search strategies were systematic, it is very difficult to present a systematic analysis of outcomes: a narrative approach to presentation has therefore been adopted.

2. National Examinations

2.1 Format of USMLE

A number of countries in the developed world, including Germany, Canada and the U.S., employ national qualifying examinations. Of these, the U.S. version (United States Medical Licensing Examination, USMLE) is probably the best studied, and will be used as an example.

The National Board of Medical Examiners was formed by the Association of American Medical Colleges in 1915, and held its first examinations in 1916 (Melnick

2006; Melnick et al 2002). Licensing of Medical Practitioners was (and is) still carried out by State Medical Boards, but success in the NBME examinations was considered an appropriate qualification for this purpose. For a number of decades, foreign medical graduates from certain recognised overseas schools could also sit the NBME examinations, and thereafter apply for state recognition. By 1956, this role was taken over by the Educational Commission for Foreign Medical Graduates, but this continued to draw on NBME materials and expertise. A variety of models of co-operation were explored over subsequent years, but over 1992-94, a joint examination, the USMLE , was implemented. Currently this consists of Step 1 ('one best answer' MCQs on science principles), generally taken after Year 2 of the medical programme; Step 2 (*Clinical Knowledge, CK*: 'one best answer' MCQs on the supervised application of medical knowledge, skills and understanding, and *Clinical Skills, CS*: an extended assessment featuring Standardised Patients and structured observed tasks), generally undertaken in year 4, and Step 3 (Medical practice without supervision: multi-answer question items and computer-based clinical scenarios), taken as the 'final' examination (USMLE Bulletin 2009). Other than Step 2 CK, the process is computer-administered. The process is currently under review once more (Comprehensive Review of USMLE, 2008). Discrimination between candidates for residency places is not one of the intended purposes of USMLE, but it is often used for that purpose.

2.2 Do national exams predict performance in later clinical practice?

Table 1 shows the results of the literature search.

If national exams are not valid and reliable predictors of performance in clinical practice, then they could be dismissed as a selection tool for Foundation. Given that this is an issue of such importance, it is surprising that it is not *the* major focus of research into assessment in countries using a national exam. However, there are a number of good studies which do relate to this issue, including an excellent meta-analysis featuring data from the USMLE. In this review, Hamdy et al (2006) conclude: *"The studies included in the review and meta-analysis provided statistically significant mild to moderate correlations between medical school assessment measurements and performance in internship and residency. Basic science grades and clinical grades can predict residency performance"*. The authors also concluded that, as might be hoped, performance on similar measurement instruments is better correlated than performance of different instruments. So NBME II scores correlate well with NBME III scores, medical school clerkship grades correlate well with supervisor rating of residents; and OSCE scores correlate well with supervisor rating of residents, when similar constructs are assessed. The results of their meta-analyses are extracted and summarised in Box A below.

Box A (NBME is the previous version of USMLE)

Predictor	Outcome	Correlation	CI	Descriptors
NBME I	supervisor rating during residency	Pearson r = 0.22	0.13-0.30	positive significant low
NBME II	supervisor rating during residency	summary correlation coefficient r = 0.27	CI 0.16-0.38	positive significant low
Clerkship Grade Point Average	supervisor rating during residency	Pearson r = 0.28	CI 0.22-0.35	positive significant low
OSCE	supervisor rating during residency	Pearson r = 0.37	CI 0.22-0.50	positive significant low
Clerkship Grade Point Average	supervisor rating during residency	Pearson r = 0.28	0.22-0.35	positive significant low
NBME I	American Board of Medical Speciality Examination	Pearson r = 0.58	0.54 – 0.62	positive significant moderate
NBME II	American Board of Medical Speciality Examination	Pearson r = 0.61	CI 0.51-0.70	positive significant moderate

Tamblyn et al (2002) compared the performance of 912 family physicians in Canadian licensing examinations with subsequent performance measured by a number of indices, such as appropriate prescribing, delivering continuity of care, and screening patients for serious illness. For instance, they noted that higher scores on drug knowledge were associated with lower rates of contraindicated prescribing (relative risk 0.88). They concluded “Scores achieved on certification examinations and licensure examinations taken at the end of medical school show a sustained relationship, over 4 to 7 years, with indices of preventive care and acute and chronic disease management in primary care practice”. This study extended and confirmed an earlier study which was confined to the first 18 months of practice (Tamblyn et al, 1998). Tamblyn et al (2007) compared performance on the Canadian Clinical Skills Examination (CSE), which is similar to USMLE Step 2 CS. Candidates who lay two standard deviations below the mean for communication skills in the CSE were significantly more likely to be the subject of non-trivial complaint in later practice. Holmboe et al (2008) explored the relationship between physicians' scores on the American Board of Internal Medicine's Maintenance of Certification examination and a variety of indices such as delivery of diabetes care, mammography and cardiovascular care. Their conclusions, like those of Tamblyn et al (2002), were stated unequivocally: "Our findings suggest that physician cognitive skills, as measured by a maintenance of certification examination, are associated with higher rates of processes of care for Medicare patients".

In all of these studies a common approach was to *categorise* performance in rank order, whether this was looking at quartiles or other fractions, or looking at those performing, say, in groupings by standard deviations from the mean, rather than by treating performance in exams or in care delivery as a continuous variable. This is probably for statistical reasons, but is fortuitously informative for standard setting purposes.

Summary Conclusion

2.2.1 There is clear evidence from a variety of sources that performance on national licensing exams is a moderate predictor of performance in later clinical practice, by a variety of measures and outcomes.

However, this is information which relates to assessment performance only, usually through cognitive knowledge measures, or less frequently through skills measures. A separate body of work has looked at the third element of Bloom's taxonomy, which in this context might be called 'professionalism'. This is addressed in Section 4.

2.3 How comparable are UK medical schools?

However, notwithstanding the conclusion in the previous section, if UK medical schools produce graduates at the same standard, then the current approach of ranking candidates within medical schools as a surrogate for ranking between medical schools can be justified.

The following theoretical arguments could be used to support the equivalence of UK medical schools.

1) The standard of admission to medical schools in the UK is relatively uniform.
2) The standard of achievement at medical school is also very high – bright students tend to continue to do well no matter what environment they are in
3) Medical schools are assessed for their equivalence through three routes.
    a. University wide support mechanisms are inspected through the QAA process.
    b. Medical courses are inspected separately through the General Medical Council's QABME process.
    c. External Examiners oversee and moderate the outcomes in all medical courses.

However, these arguments are not in themselves evidence. The literature on comparisons between medical schools is rather scanty, frequently refers to comparisons between *kinds* of courses e.g. PBL versus non-PBL, (Dauphinee and Patel, 1987; Prince et al 2003; Schmidt, Vermeulen and van der Molen, 2006), and generally relates to countries other than the UK (Verwijnen et al, 1990). However, even within these limitations, it is possible to identify some further relevant studies.

An early study (Wakeford et al 1993) indicated that there were significant differences in success in MRCGP examinations between graduates of different medical schools. This was not corrected for entrance qualification. Schuwirth et al (1999) showed small but significant advantage in problem solving for a PBL school compared to a non-PBL school. Jones et al (2002) showed differing degrees of preparedness in PBL versus non-PBL graduates, with the difference again favouring PBL graduates. Remmen et al (2001) demonstrated significant differences in basic clinical skills between four medical schools with varying curricula. Prince et al (2003), in a study of anatomy learning in schools with different curricula, found no significant differences when schools were classified by kind of curriculum, but did observe significant differences in contextual anatomy knowledge between different medical schools over all. Van der Vleuten et al (2004) compared four different medical schools on a joint test, and identified complex differences in performance, although the authors were reluctant to over interpret this data. Interestingly, they commend test equating as a strategy for situations where national testing is not present (see Section 8). Goldacre et al (2004) reported significant differences in career destinations for graduates from different medical schools. Although this does not relate to the issue of cognitive ability, it may reflect more subtle differences which are none the less important. A further study by Prince et al (2005) identified significant differences in perceived ‘preparedness for practice’ between graduates of different curricular models. A particularly well constructed study was carried out in China by Stern *et al* (2006), where significant differences were found in a sample of eight medical schools. Boursicot et al (2006) identified significant differences between pass marks (set by an Angoff procedure) and performance between five UK medical schools on a standardised OSCE. This study demonstrated that the different medical schools set significantly different pass marks for each of the OSCE stations. A replication of this study using the more reliable and valid borderline groups approach in three medical schools gave similar results (Boursicot et al 2007). This is particularly interesting, because it indicates that even setting common tests does not bring about equivalence of standard setting. McCrorie and Boursicot (2009) have reported marked differences in the assessment tools used by different UK medical schools, which seems to suggest that there will also be differences in validity and reliability in the outcome measures (see also McCrorie et al 2008).

Probably the most significant recent study is that of McManus et al (2008), because it looks at later career events. In this, the performance of UK medical graduates was studied in the three parts of the MRCP examinations with regard to the school from which they graduated. This identified statistically significant differences in performance between schools in all three parts of the examination even after correcting for admission levels (and of course, if the outcomes alone are of interest, it is desirable not to make this correction). These differences were consistent across the parts of the exam, and also showed evidence of stability in time. The Royal College of General Practitioners has also reported significant differences in pass rate when analysed by medical school.

Schmidt et al (2009) demonstrated that there were statistically significant differences in time to graduation and drop out rates between different kinds of curriculum (with ‘active learning’ style schools showing better performance). While this does not relate to cognitive knowledge scores, it does suggest that there are consistent consequences of different curriculum models.

In an innovative study looking at learning process rather than outcomes, Van der Veken et al (2009) have demonstrated that different curricula have different effects on learning styles adopted by students.

Three points should be made. First, differences in output do not imply that any medical school produces graduates below the required standard. Second, this is not a simple matter of one curricular model being better than another (e.g. PBL being better than non-PBL approaches). In the study by Prince et al 2005, the highest scores in the area under study were obtained by a traditional programme. Third, this summary does not take account of publication bias: conceivably, studies reporting a difference may be more likely to be published than studies not reporting a difference. However, on the balance of the evidence reviewed, it is reasonable to conclude that there are indeed significant differences between the standards of medical schools within any given country, including the UK. It is probably these considerations that led to the downgrading of the weighting given to the assignment to quartiles in the current Foundation selection process. It was just such considerations that have led Ricketts and Archer (2008) to argue for the introduction of a national exam.

Summary Conclusions

2.3.1 The assumption that all graduates from UK medical schools can be ranked nationally on the basis of their local results is not robustly supported by evidence

2.3.2 This does not imply that all graduates do not reach a minimum level of competence.

2.4 What about perceptions? Acceptability of a national examination

Analysis of numerical estimates of validity and reliability of assessments may lead to neglect of a significant element of the Utility equation – acceptability. By its nature, this is harder to quantify, even to estimate. But qualitative views abound, and generally seem to be of the view that national examinations have drawbacks In a recent survey by the committee evaluating the USMLE (USMLE 2008) stakeholders reported a ‘strong sentiment that the structure, timing and reporting approaches for the current USMLE make it difficult for medical schools to introduce curricular changes”. There was also a minority view that USMLE “interfered with student focus on the undergraduate curriculum”.

A GMC Consultation on assessment (GMC 2007) indentified a number of perceived disadvantages of a national curriculum (stating rather bluntly “In reality there is no prospect of a UK wide system replacing the need for medical schools to assess their own students”. These were:

- That learner centred education must allow variety between and within medical schools
- That ‘fitness to practice’ is a complex attribute, requiring a complex assessment

- That medical schools must be free to be innovative in assessment (under the assumption that a national exam removes this freedom).

However, they also indicated that “there do not appear to be significant reasons to close down policy development towards a mandatory national component to assessment”.
They also summarised a number of responses from the public consultation, expressing reservations. These included

- Preserving the variety and independence of medical schools
- Danger of over-assessing students
- Cost to learners
- Facilitating unhelpful (certainly unwelcome) league tables

In an editorial in *Medical Education,* Schuwirth (2007) summarised some arguments relating to national examinations. These were

- The ‘single shot’ (or snap shot) nature of most national assessments, even if they are on more than one occasion, versus continuous assessment of properties which may have a developmental component.

- “Limited versus extended measurements in time” where the argument (partly overlapping with the previous one) is that students in different programmes reach different points at different times. He cites one study in which significant variations were observed between medical school student rankings depending on when the test was administered, which years were compared, and which cohort of students was under investigation (Muijtens et al 2007).

- The structured nature of national assessments, in which the same or equivalent test is administered to all students, whereas local assessment may permit testing which is responsive to the individual and local needs.

- The difficulty of measuring complex constructs such as ‘professionalism’

- Centralisation of expertise. Schuwirth makes the point that national testing requires experts, while decentralised systems sometimes get away with using amateurs. He may be diplomatically making the point that he doubts if the UK possesses a sufficient number of experts in aggregate, in contrast to the Netherlands.

Ten Cate (2002) had earlier made a different but valid point about the disempowerment of teachers. “Expert teachers need to feel a sense of control over their own teaching and testing if they are to maintain an adequate degree of motivation. It is the sense of ownership and pride in education that drives teachers (*my* class, *my* students, *my* test – see how *my* students benefit from *my* guidance), rather than the impression of being a minor link in a huge educational chain. This desire has to be satisfied one way or another”

Delegates to discussion sessions organised by the MSC expressed, inter alia, the following views of weaknesses of a national assessment system

- Cost, logistics and practicality (especially with OSCE stations)
- Would undermine Medical School and GMC efforts to ensure graduates are fit to practise
- Could lead to league tables of Medical Schools
- Duplicates finals
- Would devalue extra-curricular activities
- Would encourage students to work to the exam - which could lead to the aspects of the undergraduate programme aimed at developing a professional being lost
- Poses huge problems in terms of timing given diversity in the delivery of the curriculum between different schools
- An exam can assess academic qualities but not other broad range of skills and attributes that make a good doctor.
- An exam would lack sensitivity to individual expression
- Could undermine public confidence of individual doctors of the results are publically available (Freedom of Information)

The MSC themselves expressed the following view in the Appendix to this document

- the GMC has encouraged the development of a diversity of undergraduate curricula across the UK and this innovation may be inhibited
- such an assessment would drive learning at the expense of the rest of the medical course
- results from such an assessment would inevitably be used to infer the relative performance of Medical Schools: since the available evidence indicates that the majority of the variation in performance between Schools can be accounted for by differences in admission criteria, there could be an adverse impact on widening participation due to schools focussing more on academic achievement rather than other important attributes and background of applicants
- a uniformity of curricula content would cut across health policy aimed at matching workforce to regional needs.

These views have been integrated into a summary list, in no particular order, in Box B

Box B

There is a perceived risk that a national assessment in the UK may

- endanger the variety and independence of medical schools
- create the danger of over-assessing students
- be expensive to learners (or to whoever pays for it)
- Facilitate unhelpful league tables of medical schools
- Facilitate unhelpful league tables of doctors
- Inhibit variety between and within medical schools, incompatibly with learner centred education
- Lead to neglect of the idea that ‘fitness to practice’ is a complex attribute, requiring a complex assessment
- be difficult to administer with regard to practical skills test
- inhibit medical schools innovation in assessment
- Rely on snap shot views, which are less reliable than continuous or repeated assessments
- not be responsive to individual needs
- Require expertise which may be decentralised or lacking
- distract students from valuable local assessment
- distract students from extra-curricular opportunities
- Remove a sense of ownership from teachers
- be difficult to schedule given differences in curricula
- not be sensitive to local needs

Are these views justified? Inevitably there is little evidence other than the existence of the perceptions themselves. However, in one interesting study (Wilkerson et al, 2007) University of California – Los Angeles (UCLA) School of Medicine embarked on a major curriculum reform. In addition to monitoring the Acceptability to students, performance in USMLE Step 1 was also measured. The curriculum reform was both well accepted and effective in increasing performance in USMLE against historical controls. This demonstrates that curricular change is not prevented by the existence of a national exam, and indeed the existence of national exam scores promoted the acceptance of the curricular reforms once they had been executed. However, the perception that a national exam inhibits change might prevent others from even attempting it. Acceptability is indeed a major component of the Utility Equation.

Summary Conclusions

2.4.1 Strongly held views may militate against a national examination, even if there are arguments in its favour, and its Acceptability in the UK in the current climate is likely to be low.

2.5 Test Equating

“Test Equating” means that a certain proportion of questions are in common between final examinations across different medical schools. These common questions can be used to ‘standard set’ the final examinations of each medical school, and, with comparisons over time, can be used to ensure that exams are of equivalent difficulty across time. Each medical school could then have assessments appropriate to its curriculum, but be able to compare the standard of the assessment with other schools nationally.

The selected questions then become a “High Stakes” examination in themselves and therefore have to be handled to the highest standard to avoid challenge. There would need to be a national selection body, choosing and standard setting the questions, just as with a national exam, but perhaps on a smaller scale.

Each medical school could then have assessments appropriate to its curriculum, but be able to compare the standard of the assessment with other schools nationally. It is less likely to lead to a common total curriculum. However, test equating questions could also be used as a *de facto* comparator between medical schools unless care were taken to prevent their separate release.

Summary Conclusion

2.5.1 Test equating avoids some of the problems of ‘curriculum paralysis’ which may arise from perceptions of the impact of a national testing system.

3. Interviewing Strategies

3.1 Introduction

Interviewing is a common and heavily weighted strategy in selection for medical training (Puryer and Lewis, 1981; Edwards et al 1990; Nowacek et al 1996; Patrick et al 2001; Parry et al 2006). The grey literature indicates that it is viewed as acceptable to candidates and interviewers, largely because of this familiarity, and because of its personal dimension. However, there are serious doubts about its acceptability among those responsible for selection at more senior levels, due to concerns about its subjective nature.

The search strategy identified a number of papers which are shown in Table 2. Studies were selected for discussion in the text from this table and from hand searching on the basis of their relevance to post graduate selection. More recent papers were favoured, since it is likely that the impact of social factors has changed significantly over the years, and also because more recent papers are likely to have taken previous work into account. However, some older papers are still included where they have an important and relevant message

Since many papers compare methods with each other, inevitably there will be a degree of overlap between the sections of this review. Of particular interest will be considerations of combining different selection methods to give an aggregated score (see below)

3.2 Traditional and structured interviews

One early and apparently neglected study (Murden et al 1978) showed that candidates evaluated positively by interviewers were 2-3 times more likely to receive highly positive internship evaluations. Meredith et al (1982) observed inter-rater reliability of 0.55 – 0.91 in a small sample (bear in mind, however, that inter-rater reliability is not the same as inter-occasion reliability, nor indeed validity). However, the authors did observe a correlation between interview scores and performance on clerkship assessments. Powis et al (1988) reported kappa values from 0.23 – 0.63 for two raters, while in the same year Richards et al (1988) reported inter-rater reliabilities of 0.67. However, Powis et al (1988) also report a most interesting phenomenon, which is that there was good association between performing poorly at interview and leaving medical school without completing the course. This hints at a theme which emerges from a number of sources, namely that concerns and adverse observations may be predictors of bad outcomes while ‘good’ behaviour is too diffuse a construct to predict good outcomes. Wiesner & Cronshaw (1988) in a meta-analysis of interview format and structure determined that structured interviews were approximately twice as valid as unstructured interviews, with structured individual interviews achieving a validity of 0.63 and structured group interviews achieving validity of 0.60. However they point out that there is considerable variation even between interviews of the same kind and that the confidence intervals for these estimates are very large. None the less, this tends to confirm the view that structured interviews are significantly more reliable than unstructured interviews and that they achieve acceptable reliability overall. Campion et al (1988) demonstrated that highly structured interviews based on a job analysis with anchored ratings skills and benchmarking achieved high inter-rater reliability (r = 0.88) and moderate predictive validity (uncorrected r = 0.34, corrected r = 0.56) and acceptable levels of equity and utility. These outcomes correlated with, but were better than, aptitude tests. Predictive validity was derived from studies of performance in employment. However the close match recommended between the job analysis and the interview questions is unlikely to be achievable in medical settings, as opposed to employments settings for a particular post. Edwards et al (1990) in a meta-analysis reported average inter-rater reliabilities of 0.83 (range 0.52 – 0.96), with higher reliabilities (0.83) and a smaller range of variances. Altmair et al (1992) reported that interviews of residency applicants structured around psychological properties improved significantly on the (negligible) predictive validity of the traditional interview, as measured 4 years later.

Huffcutt and Arthur (1994) adopted meta-analyses of a range of 114 studies in a variety of different fields, and calculated correlations of between 0.2 and 0.57 for interviews. However, within this range, interviews increased in validity in parallel with the degree of structure. McDaniel (1994) surveyed literature on 3 different categories of job interview. The first of these was *situational* by which they mean the interview aimed to explore responses to hypothetical situations which might arise in the course of work. The second was *job related* which explored qualifications and experiences relevant to the job in hand. The third was *psychological* and related to the determination of personal qualities. Their meta-analysis suggested that situational interviews delivered an average validity of 0.5, job related interviews 0.39 and psychological interviews 0.29. These figures were generally higher than previous assumptions about the validity and reliability of interviews had suggested. They also found that structured interviews had higher validity (0.44) than unstructured interviews (0.33). Nowacek et al (1996) described inter-rater reliabilities of 0.55

after interview. Bobko et al (1999) derived a meta-analytic matrix looking at a variety of factors and, in brief summary, concluded that cognitive ability was the best predictor of job performance, followed by structured interviews and then measures of conscientiousness. Van Susteren et al (1999) reported kappa values of 0.79, although this value may be elevated by the selection techniques used for comparison. Cortina et al (2000) indicated that structured interviews can have greater validity than measures of cognitive ability and conscientiousness ("conscientiousness" in this context means performance in a test situation rather than workplace based assessments). This was based on meta-analyses of interview performance. Luke et al (2001) observed good inter-rater reliability in a selection programme for residencies, and found both that interview scores correlated moderately with other measures, but also explained a significant proportion of the variance in admission. Posthuma et al (2002) provide a comprehensive narrative review of factors which can influence performance in interview. Although this approach does not lead to quantitative estimates of interview performance, it is extremely valuable in partitioning components of possible interactions in such areas as disability and prejudice. Khongphattanayothin et al (2002) found a moderate to strong correlation (0.69) between scores on the Cumulative Grade Point average of applicants and the Paediatric In-Service Examination, while letters of recommendation and interview scores were not predictive. However, interview score correlated moderately with performance on their Clinical Performance Rating Scale (CPRS), and there was a weak non significant correlation between letters of recommendation and CPRS.

While much of the above evidence is supportive of the belief that structured interviews can reach moderate levels of reliability, a study by Kreiter et al (2004) indicated that there is a high degree of context specificity, and observed low to moderate reliability only. Their generalisability study revealed a strong context specificity effect – the variance due to the interaction between candidates and occasions was high. Olawaiye et al (2006) found a significant correlation(0.60) between rank on a structured interview process and a subsequent clinical performance score in residency. Mallott (2006) correctly points out that a significant problem is distinguishing between the process of maturation and "significant psychopathology" although psychopathology may be pitching it a little strong. Randall et al (2006) compared performance on a range of different selection methods within an assessment centre process. Candidates rated the process as acceptable, and interview gave results related but not identical to the other selection methods used.
Brindal and Goodyear (2007) found a significant correlation between interviewer scores and performance three months into SHO posts. Smith et al (2006) used a multi-station approach similar to MMIs (see below) in interviewing neurology SpRs. They found that traditional interviews corresponded less well to the overall ranking than any other method. Goodyear et al (2007) used a three station interview and a structured short listing approach to select for paediatric SHO posts, and found high values of G (Generalisability Theory's summary of reliability) of at least 0.8.

Rao (2007) described an extended test of the reliability of a particular structured interview process in selection SHOs for psychiatric posts. They concluded that the inter-rater reliability was high, although no conclusions could be drawn as to validity in this study. Hamel et al (2007) reviewed candidates in line with the CANMEDS competencies, and demonstrated that interviewing in four teams of two interviewers gave good reliability. However, Thordardson et al (2007) compared the performance of orthopaedic residents on interview score at entry or USMLE Part 1 scores with their subsequent performance in professional exams taken later in their training, and found only fair or poor correlations. Gallagher et al (2008) showed good concordance between a range of measures used to select candidates for higher surgical training, including simulations, interviews and an assessment of their suitability for a career in surgery. Westwood et al (2008) also found that candidates regarded structured interviews as acceptable as a means of selection. As Albanese (2004) indicates, an interview may indicate that “a school values the personal interaction between human beings”. This personal and human dimension may in some circumstances be as important as technical considerations of validity and reliability.

Summary Conclusions

3.2.1 A considerable body of evidence indicates that traditional interviews have low reliability and validity

3.2.2 A considerable body of evidence suggests that structured interviews have moderate reliability and some evidence of predictive validity.

3.3 Legal consequences

A different approach was taken by Terpstra et al (1990), in which they studied litigation arising from a variety of selection methods. They found that unstructured interviews were the most likely to lead to subsequent litigation, followed by tests of cognitive ability and tests of physical ability. Conversely, structured interviews, work samples, assessment centres and personality tests were significantly under-represented in litigation as compared to their frequency of use. The authors also considered (on much smaller samples) the outcome of litigation, and found that the litigant was successful most often in the case of unstructured interviews, followed by physical ability tests, cognitive ability tests and work samples. Structured interviews and assessment centres survived all of the challenges that had been mounted against them (admittedly a small number). Similar conclusions were drawn by Posthuma et al (2002), who concluded that structured interviews were defensible in law.

3.4 Factors affecting interview performance

An interesting study by Boor et al (1983) found evidence that males and females were rated differently at interview, with female candidates being judged more on the basis of their appearance and demeanour, whereas with male candidates appearance was not important, although demeanour remained a significant factor. Smilen et al (2001) found that knowledge of previous cognitive knowledge test scores (such as USMLE) biased the results of the interview process, with interviewers showing concordance in their awarded interview scores with known previous test performance. Swanson et al (2006) found the same effect when comparing ‘blinded’ and ‘unblinded’ interviews with respect to USMLE scores, and also concluded that a blinded approach should be adopted.

3.5 Multiple Mini Interviews

Eva et al (2004a) have described an approach to interviewing for undergraduate places derived from the Objective Structured Clinical Examination (OSCE). This approach, known as Multiple Mini Interviews (MMIs), takes applicants through a series of ten short highly structured interviews focussing on a single aspect of the desired features for candidates. Reliability in the initial study was observed to be 0.65, and “the variance component attributable to candidate-station interaction was greater than that attributable to candidates”. Acceptability was good for both candidates and interviewers. A follow up study of the participants (Eva et al 2004b) offered some evidence of predictive validity from this approach. Compared to interviews, personal statements and grade point averages, the MMI was the best predictor of OSCE performance (Beta = 0.44) while the grade point average was the best predictor of cognitive knowledge test performance. Even deliberate security violations of the MMI structure did not influence the outcomes (Reiter et al 2006) An examination of rater characteristics hinted at the value of having heterogeneous interviewers, but again returned high values for reliability (Eva et al 2004c). Continuing the follow up of the study cohort of students, Reiter et al (2007) were able to report that MMIs were the only significant predictor of clerkship performance, compared to measures of cognitive ability and other purported measures of non-cognitive performance. Rosenfeld et al (2008) provide a recent review in which they summarise the data showing that MMIs are more reliable and have higher predictive validity than traditional interviews, and also consider the cost of administering an MMI programme, concluding that those are comparable with costs of traditional interviewing methods. Dodson et al (2009) have added to evidence on costs by demonstrating that 5 minute interviews are as reliable as 8 minute interviews. Hofmeister et al (2008) explore the acceptability of MMI for international medical graduates in Alberta Canada, and found high acceptability for both interviewers and candidates.

David Powis has indicated (personal communication) that validity is increased when the interview structured to obtain direct evidence for a quality, and reduced when an interviewer has to draw inferences from unverifiable statements. Examples of indirect assessment where inference is required are such subjects as motivation for medicine, teamwork capacity, coping strategies, ethical orientation, conscientiousness, reliability, attitudes and values, and potential compatibility with the program. Examples of qualities that can be assessed directly at interview are verbal communication skills, interpersonal skills, reasoning skills, lateral thinking, problem solving, quality of reasoning, creativity, decision making, organisational skills and capacity for remaining calm under pressure. To quote Powis directly: “It’s not about what they say….but about what they do”. It may well be that the strength of MMIs lies in the fact that they often appear to be directed towards verifiable qualities that can be directly observed.

Summary Conclusions

3.5.3 There is strong and developing evidence that MMI approaches have stronger reliability and predictive validity than other interviewing techniques.

4. Non-Cognitive Properties

In addition to assessing cognitive abilities, it is also important to assess non-cognitive abilities, only a few of which are explored by skill tests such as OSCEs.

4.1 Professionalism and the Role of Conscientiousness

While there are good tools available for measuring cognitive knowledge and good tools available for measuring clinical skills, tools for measuring professionalism are poor, (Lynch et al, 2004; Shrank et al 2004; Stern 2006; Jha et al 2007; Thistlethwaite and Spencer 2008). This is unfortunate, since it may be the most important of the three domains: analyses consistently reveal that lack of professionalism is a significant factor in disciplinary proceedings, even compared to knowledge and skill.

There has been gathering evidence that adverse behaviour at medical school has predictive power for subsequent difficulties in clinical practice. Wright and Tanner (2002) demonstrated that failure to provide a passport photograph on entry was statistically significantly correlated with the likelihood of subsequently failing final examinations. Although the outcome measure was still under the aegis of the medical school, it would normally be expected that students sitting finals were close to achieving appropriate professionalism.

In what are now regarded as seminal studies, Papadakis et al (2004, 2005) showed that analysis of Dean's Letters and medical records for 'negative' or ambiguous statements showed a correlation with subsequent state medical board disciplinary action compared to a control group in a retrospective case-control study. Stern et al (2005) explored a variety of parameters gathered at admission, failure to complete evaluations, immunisation compliance, and self assessment accuracy as predictors, and compared them with subsequent review board problems and clerkship evaluations. They found no significant correlations with admissions data, but did find significant correlations with all of the others.

Ainsworth and Szauter (2006) studied 'Early Concern Notes' to identify themes which caused concern to staff. These were:

- Professional responsibility/integrity (failures of reliability)
- Pursuit of excellence (doing the minimum)
- Personal interactions

A parallel study was conducted of themes that emerged from state disciplinary events, and it was concluded that the same themes emerged from the relevant hearings and documentation, indicating that the same factors as caused concern at medical school underlay later disciplinary problems.

Papadakis et al (2008) found that poor performance on behavioural and cognitive measures during residency was associated with greater risk for state licensing board actions against practicing physicians at every point on the performance continuum. Programme Directors had graded residents on an anchored Likert scale, of which the lowest point on the 'Professionalism' axis was:

"lacks altruism, accountability, integrity, commitment to excellence, duty, service, honor; disrespectful to other health care professionals; irresponsible; unreliable; not punctual; ineffective communicator; disruptive; disorganized; records tardy and/or illegible."

These findings were taken as support for the Accreditation Council for Graduate Medical Education standards for professionalism and cognitive performance, and as pointing the way to the development of best practice to remediate such deficiencies.

Interestingly, this paper also reported an association between academic performance and subsequent disciplinary action. This was summarised by Maxine Papadakis (personal communication) as "It's good to be good *and* good to be smart". However, many of the characteristics also involve what might be called diligence or conscientiousness.

Three meta-analyses (Barrick & Mount, 1991; Mount & Barrick, 1995; Tett, Jackson & Rothstein, 1991) of the relationships between the Big Five personality factors (Thurstone 1934) and job performance have found conscientiousness to have a weak to moderate relationship with job performance across a variety of jobs and settings. Barrick and Mount (1991) found the overall true score correlation (i.e., corrected for range restriction and unreliability in the predictor and criterion) between conscientiousness (as derived from facets of conscientiousness, cf. Mount & Barrick, 1995) and performance to be approximately 0.22, although Tett et al (1991) found the relationship to be 0.179. Mount & Barrick (1995) found an uncorrected correlation of 0 .18 which, when corrected for criterion unreliability, range restriction, and facet intercorrelation, adjusted to 0.267.

There are, of course, personality factors other than conscientiousness. Nevertheless, in this review, it is worth focussing on conscientiousness for two reasons. First, in multiple studies, conscientiousness has demonstrated a significant relationship with performance in most jobs (Barrick & Mount, 1991; Hough, Eaton, Dunnette, Kamp & McCloy, 1990), although other dimensions of personality, such as extraversion, have proven to be significant predictors of performance for some jobs but not others (Barrick & Mount 1991). Second, results from different meta-analyses are consistent with respect to the validity of Conscientiousness (e.g., Barrick & Mount 1991; Mount & Barrick 1995). Due in large part to these findings, much research in the area of personality has focused on conscientiousness, with considerably less interest in the other personality dimensions (Barrick, Mount & Strauss, 1993; Schmidt et al., 1992; Stewart, Carson, & Cardy, 1996). In addition, the work of Papadakis and colleagues clearly suggests a major role for conscientiousness in predicting later performance in clinical practice.

Building on these considerations, the lead author has recently developed an objective scalar measure of conscientiousness among medical students (The Conscientiousness Index, CI: McLachlan et al, 2009). This is based on tabulating all occasions on which students might be conscientious or diligent (e.g. in attending compulsory sessions, completing course evaluations, submitting work on time, completing essential documentation such as placement allocation forms etc. Where students meet the requirements, they are awarded a CI point but not otherwise. Reliability is moderate to good (Split year and year on year correlations of approximately 0.6. Validity was explored by comparing with independent staff estimates of professionalism. These showed strongly significant associations, particularly at the low end i.e. students who scored low on the CI were also those evaluated as causing concerns about lack of professionalism. The CI is easy and inexpensive to collect, and offers a potential method of ranking students in terms of their behaviours, observed on many occasions throughout the year. Interest has been expressed in adopting the CI in other undergraduate settings, but also in post graduate settings.

Another approach has been through the use of Critical Incident forms, especially since this approach has been shown to be valuable in professional fields such as anaesthesia (Rhoton 1989; Rhoton et al 1991; Rhoton 1994). Here the issues are the continuity and consistency of the record keeping system, and the ease with which it can be summarised and transferred to subsequent professional environments.

A traditional method in the US of representing such information has been the Dean's Letter. This has suffered from lack of consistency in record keeping, and subjectivity both in recording information and in presenting it. However, the work of Papadakis et al has shown that it none the less has predictive value.

4.1.1 Behaviour which causes concern at undergraduate level has predictive validity for later clinical practice.

*4.1.2 Conscientiousness is a significant component of such concerns, and may be possible to measure as an objective scalar property**

4.1.3 Critical Incident forms capture valuable information about adverse events which are below the threshold of referrals to Fitness to Practice Panels Panels.

4.1.4 A narrative summary of information relating to behaviours in Undergraduate programmes would be of value for selection decisions for Foundation places.

4.2 Personal Qualities Assessment

Considerable interest has been shown in measures of personal qualities, by which is meant non-cognitive traits measureable under test conditions. Albanese et al (2003) have identified 87 personal qualities in the generic sense which may be relevant to the practice of medicine, so plainly the task is not a simple one. A variety of instruments have been developed and demonstrated reliability – for instance the Jefferson Scale of Physician Empathy (Hojat et al 2001). Powis and colleagues (Powis et al, 2005; Munro et al 2005) has described use of a scale measuring narcissim, aloofness, self-confidence and empathy (NACE), as a potential tool for selection, and have explored (Bore et al 2005) a measure of moral orientation (Mojac) which can be used as a separate dimension for placing students in a two dimensional space. These instruments, however, have not yet been shown to have predictive validity for excellent clinical practice, and cannot therefore yet be recommended at this stage as a high stakes selection tool.

This specific conclusion is reinforced by the conclusion of Hemmerdinger et al (2007) who carried out a review of 59 potential instruments of empathy measure. They concluded “no empathy measures were found with sufficient evidence of predictive validity for use as selection methods”.

As suggested by the work of Powis et al (1998 – see Section 3) there is at least the possibility of negative selection – of choosing those (presumably few) candidates not suitable for medical training at some level. One instrument (the Hogan Development Survey – Knights and Kennedy 2006) has been designed with the intention of identifying those negative traits which may not become apparent at interview. However, this study did not explore correlations between dysfunctional scores and performance in medical settings.

With regard to ‘negative selection’ Ferguson et al (2003) report the existence of evidence that negative referees’ reports have predictive value for future performance.

Summary Conclusions

4.2.1 Personal Qualities Assessment instruments currently lack the evidence in medical selection which would permit their use for high stakes selection.

5. Use of records of personal achievement

Results of the literature search are provided in Table 3.
Records of personal achievement, or portfolios, are now a standard part of many medical assessment programmes. They may be described as a collection of information intended to demonstrate the achievement of trainees, and may include some or all of the following: an encounter log; a record of educational experiences; a description of critical incidents; details of performance scores on assessments; and examples of reflection on performance and learning. Despite the frequency of their use, there is surprisingly little evidence on their predictive validity. Our literature search indentified some 30 relevant papers, none of which addressed predictive validity, and only a few discussed other meanings of validity. One recent review by assessment experts (Driessen et al 2007) focussed more on the process than on measureable outcomes. Portfolios also present significant problems in the process of assessment, both in the subjective nature of responses, and in the time cost involved One paper in which the assessment process was described in detail required 170 examiner-minutes per candidate consisting of 60 minutes of portfolio reading by two different examiners (= 120 minutes examination time) and two 25-minute oral examinations conducted by different examiners (= 50 minutes examination time) (Davis et al 2001).

However, some evidence of reliability can be found, although Pitts et al (1999) concluded that summative judgements could not be made safely in their setting. Gadbury et al (2003) found good internal consistency, and predicted the number of subscales and raters (three raters) that would be necessary for good reliability. Driessen et al (2005) found that inter-rater reliability ranged from 0.46 to 0.87.

Melville et al (2005) in a study of paediatric specialist registrars, found moderate inter-rater correlations (0.52), and, using the same decision analysis approach as Gadbury et al (2003), concluded that four raters would be needed for good reliability. O'Sullivan et al (2004) estimated that three raters would be required for criterion referenced decisions. Rees and Sheard (2004) found two raters could achieve total inter-rater reliability of 0.771, although agreement on individual items was markedly lower. However, these studies all describe systems in which a great deal of time and effort is spent on developing the test methods and training skilled assessors. While individual medical schools are likely to be devoting such efforts to their own portfolio based systems, these vary so much that no relative ranking could be arrived at between systems, and there is unlikely to be support for introduction of a UK wide system of portfolio design and assessment, given that portfolios may well represent reflections of the unique nature of each course. If three or more trained raters are likely to be required, then the time costs compared to other assessment methods are likely to be prohibitive.

Of course, there are also time costs for those being assessed, and this may affect the acceptability of this approach. Davis et al (2009) reported rather mixed feelings on the part of undergraduate medical students. Hrisos et al (2008) also identified reservations among Foundation doctors, with the telling descriptor "*burden*" being used. Gordon (2003) did find high acceptability for portfolios and interview, but this was in an unusual situation where this was the sole method of Year 1 assessment for medical students. Dornan et al (2002) found significant reservation expressed among endocrinologists, as well as positive views, a common finding which seems to indicate that views on portfolios may often be polarised.

In general, the caution recommended by Roberts et al (2002) in using portfolios for high stakes testing remains appropriate.

Summary Conclusions

5.1 Despite their wide spread use, the predictive validity of portfolios is uncertain

5.2 Under defined conditions of design and training, portfolio assessment may be moderately reliable.

5.3 The cost of administering a portfolio based assessment system comparable across medical schools is likely to be high.

5.4 The general acceptability of portfolio based summative assessment to candidates appears to be low

6. Use of Autobiographical Statements

Too few papers were found by the literature search to be worth tabulating.
There is essentially no evidence on the use of ‘white space’ information in the Foundation selection process. Extrapolations must be made from evidence on other forms of personal statements.

McManus et el (1986) found a slight negative relationship between analyses of applicants’ statements and performance outcomes.

Gurza-Dully and Melaney M (1992) explored the predictive validity of a number of selection methods with a population of respiratory therapists. Other than the neatness of the forms, there was no *significant* positive predictive values of application form entries, but low positive values were recorded for some other aspects.

Lamsdale et al (1999) explored the use of structured and competency based questions for applicants to the police force. In trials on serving police officers, they found that there was a correlation between responses and performance, and when used with actual applicants, they observed a significant correlation between their question score and their subsequent performance.

Ferguson et al (2000) found no evidence that the content of candidates personal statements predicted success in the early years on an undergraduate medical programme. Ferguson et al (2003) found that the presence of more information in the personal statements was predictive of performance in approximately one third of the subsequent assessments used in medical school, but this was a low value compared to other selection methods.

Patterson et al (2009) reviewed a variety of selection methods for post graduate training, including the use of structured application form questions provoking open ended answers, not dissimilar to those used in Foundation selection. These were found to have good reliability (Cronbach’s alpha = 0.78). The corrected correlation with performance at a selection centre was 0.40, suggesting moderate predictive validity. The authors conclude that this method is a moderately successful predictor of selection purposes, and adds incremental value to tests of knowledge. However, they also comment on the significant costs of scoring open ended questions, and identify merits in using situational judgement tests (McDaniel et al 2001).

Summary Conclusion

6.1 Personal statements made in response to structured questions have moderate predictive validity for subsequent performance, but are expensive to score.

7. Combining Selection Strategies

Much attention has been given to individual selection methods: much less has been given to the impact of combining methods to produce an aggregate outcome, which potentially has important confounding effects. Two factors are important (McLachlan and Whiten 2000; Eva and Reiter 2004). The first of these is the variance of the outcomes. Combining two measures with different variances unwittingly weights the outcome towards the measure with the largest variance. Scores should be converted to standardised distributions such as z scores before combining. The second is the reliability of the outcomes: an aggregated measure should be weighted in accordance with the reliability of the individual outcomes. Eva and Reiter (2004) demonstrate clearly through examples the adverse consequences of failing to take these factors into account.

A separate issue is that of compensation between domains or elements. If more than one domain is considered (e.g. knowledge, practical skills and behaviours) is a single aggregated figure to be produced or are domains to be non-compensatory, requiring at least defined minimum performance on each?

Summary Conclusions

7.1 Where different selection methods are combined, scores should be converted to a standardised distribution and weighted according to their reliability.

8. General Discussion, Conclusions and Recommendations

Of the current methods of selection, ranking based on the equivalence of medical schools probably has some *defensibility* (see Appendix for more on this term), but there are also some challenges. The existing evidence does not strongly support the idea that UK medical schools are equivalent. There is also a potential argument that a single aggregated measure cannot capture all aspects of student performance: many medical schools attempt to measure skills, knowledge and behaviours separately, and there is little to commend putting these together into a single ranked figure. Then, too, this approach explicitly requires the use of assessments generally designed to be competence measures as discriminator measures, which is widely seen as inappropriate. Furthermore, for the next few years at least, there is an issue of lack of transparency. Medical students were encouraged to view their studies as competence measures, not discriminator measures, and encouraged to co-operate with each other and to spend time on interests not directly related to the traditional curriculum, through SSCs and other means. Retrospectively imposing a competitive element cannot be seen to be good practice. An interesting report from Liverpool suggests that students may be declining to participate in peer assessment exercises because of this new competitive ethos (Garner and O’Sullivan, 2009).

The scoring of ‘white space’ questions has low to moderate validity on its own. However, little evidence is available on the reliability of this process as applied in Foundation selection. Confidential information the lead author has seen suggests that reliability may be an issue in some Schools. Acceptability may be an increasing issue – where self-reports are used there are always issues of authenticity of statements, but also of the perceived authenticity of statements.

However, the aggregation of these two elements does not have a defensible basis that we know of. The calculation of how many points are to be allocated to the ranking and how many to the white space questions ought to be based on the variance and reliability of each of these methods (See Section 7). If this has been done, it has not been published.

In summary, there is a risk that the current method might be seen as 'capricious' in Norcini's terms.

Certainly there are technically better alternatives. In terms of cognitive knowledge and skills testing, there is evidence that a national examination process like the USMLE would be a predictor of clinical performance at Foundation with moderate to good validity and reliability. However, the utility of this approach would be compromised by concerns about its cost, its educational impact, and its acceptability (particularly to medical schools). In terms of cost, a national exam would require a National Exam Board as in the US, to set the examinable curriculum, commission, field-test and select questions according to a blueprint, administer the exam nationally and issue the results in a diligent way. Probably only the GMC has the resources to do this. There would also be problems of timing, given the widely different nature of medical curricula. Its educational impact, as the NBME survey showed, lies in the fear that students would study only for the test, rather than to deepen their understanding of medicine. However, acceptability is likely to be the greatest problem. It is clear from the views gathered by the MSC itself that a national exam process would be regarded with grave reservations by a number of stakeholders, including heads of medical schools concerned about the development of an inappropriate 'league table' of schools, and students concerned about the quality of their education in general.

However, test equating of assessments of cognitive knowledge may help resolve at least some of these issues. In this approach, a proportion of selected response questions of known facility and discriminating power would be drawn from a national question bank, and employed in what each medical school considers to be its last test of knowledge. This common pool of questions can then be used to draw an inference about the level of the rest of the examination, and to permit a calculation of the relative performance of candidates across all the participating schools. This is essentially the approach taken by educational testing organisation such as Pearson Vue and the Australian Council for Educational Research in ensuring consistency of standards from test to test and from year to year.

Two plausible question banks already exist – the Universities Medical Assessment Partnership (UMAP) and the General Medical Council's Professional and Linguistic Assessments Board (PLAB) Part 1 banks. Other educational providers may be able to bid to provide a service along these lines.

Characterisation of the properties of assessment items allows the use of Computer Adaptive Testing (CAT). In this approach, candidates follow a path through the assessment which is modified by their performance. In other words, if you get the first question right, you get a harder one, if you get it wrong you get an easier one. This reaches reliable estimates of candidates' ability (not 'knowledge' since knowledge has a case-specific component) more quickly than in a standard test. This might be a possible approach in the future.

USMLE now incorporates a directly observed skills test, although the format of this is currently under review (USMLE 2008). Test equating is much more challenging for this part of the process. All medical schools have similar tests of knowledge, using selected response items. But UK medical schools employ many variants of the OSCE and other skills test, particularly in later years. Although PLAB does have a bank of validated OSCEs, it might be difficult to equate the PLAB OSCEs from their bank with the existing skills tests. It would be desirable to test this through pilot studies similar to those conducted by Boursicot et al (2006, 2007) in advance of their use nationally.

An alternative is the use of a ‘selection centre’ approach for skills. ‘Selection centres’ represent an approach rather than a location. Their use is fully summarised in a recent text (Patterson and Ferguson, 2007), which also sets their use in context with a variety of selection methods. Briefly, it employs a multiple look, multiple method approach, in which the methods are designed around the job specification. The use of a number of observers acting independently improves the reliability of the results as demonstrated by Generalisability Theory. It has been shown to have good predictive validity in medical settings (Randall et al 2006 a-c). Patterson and colleagues refer to the need for a thorough analysis of the job requirements before setting up a multi-method testing system.

It might be thought that this would merely create a national OSCE, with all of the attendant difficulties. However, some Foundation Schools are already introducing OSCE style tests of FY1 doctors, in order to establish their individual skills competence, and this sometimes extends to ‘stand alone’ FY2 applicants who have not undertaken FY1 before applying. If the Foundation School were responsible for administering the regional selection centre, this would break the link between medical schools and candidates, since there may be more than one medical school per Foundation School region, and there would be no need to categorise the results by medical school.

Structured interviews have moderate predictive validity and reliability. More importantly in the real world, they have a high degree of acceptability. Candidates may value the chance to present themselves as a person, rather than as an application, and structured interviews are less likely to lead to successful legal challenge than some other methods. The approach adopted by Kevin Eva and colleagues, of applying the principles that make OSCEs successful (multiple looks by multiple observers) to the interview process through the MMI currently seems to offer better predictive validity and reliability even than structured interviews. It is easy to imagine an MMI, designed around a careful analysis of the required and directly observable qualities, being incorporated into a selection centre approach.

However, we must consider an even more serious issue, currently not well explored in terms of selection procedures. Most difficulties in later practice do not arise through deficient knowledge or the deployment of inadequate technical skills. They reflect behavioural or attitudinal issues – what might generically be called professionalism. A well designed selection procedure would include this property also.

The current Transfer of Information Process does not meet the legitimate need of Foundation Schools to access relevant information, since it is properly designed to support the student. Equally the blunt declaration required of medical school Deans that a student is fit to practice (or not) is not helpful, since any declaration that a student was not fit to practice would lead to legal challenge unless there had been a successful Fitness to Practice procedure – in which case it would not be necessary.

Papadakis and colleagues have provided evidence (summarised above) that concerns at undergraduate level are low to moderate predictors of concerns in later clinical practice. This property appears sensitive rather than specific – many of whom concerns are expressed do not encounter significant difficulties. However, there have been problems in collecting the relevant data in a consistent manner, in overcoming the subjective nature of many of the observations, and in transferring information forward. Fitness to Practice procedures are of such severity and complexity that they are rarely invoked, and do not capture low level on going matters of concern. One option would be to employ Critical Incident Report forms, as is done in the NHS itself. Many medical schools already do this, and it is possible to imagine the design of a form that commands nationwide support, particularly if it incorporates space for student response and reflection. This could be associated with a Dean's Report at the end of the undergraduate period, again in a common format, which allowed Deans to comment in a structured way on issues that might have arisen during each student's career. Foundation Schools could then analyse this document through reviewers as part of their deliberation process.

Another approach might be through the collection of continuous objective data reflecting conscientiousness. (Here the lead author declares an interest as the author of a paper in press on this topic: McLachlan et al 2009). As described above, this records objectively each occasion on which an applicant might choose to show conscientiousness or not, and tabulates these as a Conscientiousness Index. This approach shows concurrent validity with other estimates of professionalism, and is easy and inexpensive to administer, and further is collected continuously throughout the student's undergraduate career, rather than being based on occasional observations which the student can fake.

Together, these considerations lead to a number of recommendations.

Recommendations

An evidence based, acceptable highly defensible, selection approach for Foundation places could therefore take the following form.

1. Test equating of cognitive knowledge (CK) assessments in each medical school would give candidates a national 'Knowledge' rank equivalent.

2. Skills assessment (either through test centres administering a common OSCE programme) or, more problematically, through test equating of OSCE skills test within existing curricula) would give each candidate a 'Skills' score.

3. Each Medical School Dean would present a structured national report form of 'behaviours' for each candidate for review by the Foundation School, based on consistently collected information.

4. Each candidate would receive a structured interview (either as a single interview or through Multiple Mini Interviews) designed to explore those qualities relevant to the job which are amenable to direct observation (e.g. such as interpersonal skills, verbal

communication skills, and problem solving). This would generate a national ‘Interview’ score.

5. Foundation Schools could weight these scores according to defined, justifiable and transparent criteria in making selection decisions.

Appendix 1 Background and terminology

1.1 Domains of Assessment

A standard taxonomy (Bloom 1956) is that there are three major domains that should be assessed: Declarative Knowledge (perhaps a combination of ‘reasoning ability’ and ‘learning’), Procedural Skills, and Professional Behaviour. There are good assessment tools for the first of these, reasonable tools for the second, and few validated tools for the third.

1.2 Purposes – Competency and Discrimination.

Assessments can be intended either to assess competence (‘do all candidates meet a minimum standard?’) or to discriminate between candidates (‘where do candidates fall with respect to each other on a particular scale?’). Each assessment should be designed for its purpose. For instance, a competence assessment should be most sensitive at the borderline between pass and fail. Discriminator assessments, by contrast, may be designed to be most sensitive in the middle of the range, where most candidates are found. And, naturally, the scoring and reporting scales are different for each kind of assessment. For competence assessments, only two scale points are required – pass/fail, competent/not competent, both for individual assessment items, and for the assessment as a whole. For discriminator assessments, many more points are necessary, and the fineness of the scale required relates to the number of candidates and the intended purposes of the discrimination. Competency Assessments require Criterion Referencing approaches, while Discriminator Assessments benefit from Norm Referencing (see 1.4 below).

1.3 Purposes – Formative and Summative

Similarly, the distinction between formative and summative purposes is well known – formative assessments offer feedback to candidates and summative assessments determine progression. A widely agreed assessment principle is that formative and summative tests should be kept separate. For instance, Stern (2006) says “*Evaluators must decide the purpose of evaluation prior to developing an evaluation system…Educators planning both formative and summative assessments should use separate and independent systems*”. However, all summative assessments can have formative consequences.

1.4 Standard setting – norm referenced and criterion referenced

Norm referenced standards are established with regard to a reference population, while criterion referencing relates to absolute standards. Although most current standard setting methods rely on criterion referencing, it is important to consider the purpose of the assessment – competency purposes require criterion referencing while norm referencing is best for discriminator measures. It is not widely appreciated that norm referencing is generally more reliable than criterion referencing precisely because it does not require the setting of an absolute standard. For instance, I cannot tell you against a standard how happy I am from day to day, but I can say with some reliability that I am happier on some days (Saturdays) than others (Mondays, for example).

Criterion referenced standard setting may focus on the properties of the test item (e.g. Angoff and Ebel methods), or on the properties of the test takers (Contrasting Groups and Up and Down methods).

1.5 High stakes examinations

When important consequences arise from an assessment, it is generally described as ‘high stakes’. Summative assessments in medicine are almost by definition high stakes, and this also applies to selection for F1, F2 and training posts. A national exam would be higher stakes than one confined to one medical school.

A high stakes exam should be clearly defined as to purpose. It should be ‘blue printed’ i.e. matched against a curriculum which must itself be defined in advance. The development of assessment items requires assessors to be trained, benchmarked and audited. Assessment items should be field tested, and there should be a feedback loop which allows for performance (see below) to be evaluated. The size of the assessment must be suitable to the task. Appropriate standard setting methods must be employed, involving expert staff. Storage and delivery of the assessment items must be secure.

To deliver a national level high stakes exam, an organisation capable of obtaining, testing and administering the equivalence questions in a professional, competent and confidential way would need to be established. This would require selection, training, benchmarking and auditing of question setters. It would be necessary to create a question bank in which performance details of questions was recorded, and to select questions from the bank by means of a blueprint. Since questions would have to be sent to a variety of environments, secure means of communication would have to be established.

1.6 Item Performance

Assessment items can be more or less easy. This property is called *Facility*. If the question is too easy, then most candidates can answer it correctly (high facility). Conversely, if a question is too difficult, few students can answer it (low facility). The *Discrimination* of a question shows the range of responses it receives. It might be helpful to think of discrimination as being like the standard deviation of the distribution of the answers, while facility is in some ways like the mean. Finally, a question may be answered correctly by weak students and incorrectly by strong students. This can be thought of as a correlation (and for MCQs, is calculated as the *Point Biserial*).

A sophisticated way of looking at the performance of each individual assessment item is *Item Response Theory*. This approach is used by professional testing organisations, such as the Australian Counsel for Educational Research (ACER) and the National Board of Medical Examiners (NBME) in the USA.

Once the performance of individual items has been determined, these can be combined in various ways according to the purpose of the assessment. For instance, a competency assessment can be designed to be most sensitive in the pass-fail zone, while a discriminator assessment might combine items with a much wider range of facilities and strong discrimination properties.

1.7 Utility

Utility was helpfully summarised by Cees van den Vleuten as

Utility = V x R x E x A x C
where
V = Validity
R = Reliability
E = Educational Impact
A = Acceptability
C = Cost

However, this might better be described as a general relationship than an equation, and the construct of Defensibility (capable of withstanding professional or legal challenge) should be added. Hence, a better formulation is:

Utility is a function of Validity, Reliability, Educational Impact, Acceptability, Cost and Defensibility.

1.8 Validity

Overall, Validity is the degree to which a test measures what it is intended to measure. It relates to Reliability in somewhat complex ways — a measure with low Reliability is sometimes described as being excluded from having high Validity - but Reliability and Validity cannot be traded off against one another in a simple way as is sometimes assumed.

There are a variety of sub-types of validity. Their meanings may sometimes be controversial, but the following operational definitions are used here.

Face Validity: Whether an item makes sense to a panel of experts. One can usefully ask this of one item or question.

Content Validity: Whether the items in an assessment accurately represent the domain being tested e.g. fair sampling. One can usefully ask this of one test or group of items.

Criterion Validity: Drawing inferences between scale scores and some other measure of the same construct. One can usefully ask this of one or more tests. There are 2 sub-varieties of criterion validity. Concurrent Validity is when correlation of one measurement is observed against another measure of known or supposed validity at the same time. Predictive Validity is when correlation of one measurement is observed against another measure of known or supposed validity at a future time.

Construct Validity: A test of the underlying construct. One can usefully ask this of one or more tests. This is the hardest to understand, but an example of a construct is that in a test, higher scores will be progressively obtained by those with increasing levels of expertise. So a test of construct validity would be to give a medical performance test to 1st year students, 5th Year students, Foundation Year 2 doctors, registrars and consultants.

Convergent Construct Validity should be positive where tests are assumed to measure the same construct and Divergent Construct Validity should be negative where tests are assumed to measure different constructs.

1.9 Reliability

Reliability is the degree to which an assessment measures with consistency.
There are several different ways of approaching this.

In Classical Test Theory (also known as Classical Measurement Theory, 'True Score' Theory), it is assumed that any given Score consists of a True Score plus an Error. The error is treated as being of one kind, and it is assumed that the Error can be estimated. Typical tools for exploring this kind of error are Test-Retest estimates, Cronbach's Alpha and tests of inter-rater reliability such as Kappa.

In Generalisability Theory, errors are treated as arising from a number of sources, each of which can be explored and measured separately. More technically, it considers all sources of error (factors) and their interactions, e.g. candidate, marker, item, student-with-item, marker-with item, marker-with-student, and marker-with-student-with-item. A Decision analysis (D study) can be carried out with this information, which allows prediction (for example) of how many raters are required in an assessment.

In Item Response Theory, the underlying construct is that there is a relationship between the probability of a candidate answering the question correctly, and the ability of the student. This is expressed as the Item Characteristic Curve. This sophisticated, powerful but complex interpretation is widely but probably exclusively used in national and large commercial testing organisations.

1.10 Educational Impact

This describes the impact the assessment strategy has on the learner's learning strategies. It may be positive in aligning the learner with the goals of the programme, or it may be negative, as in promoting a 'cram and forget' strategy, a focus on trivia, or an unhealthy form of competition rather than co-operation.

1.11 Acceptability

This describes basically whether or not the learners and assessors *like* the assessment approach. This is often coloured by familiarity, and sometimes by engagement with a personal dimension, as in interviewing, rather than by more abstract qualities such as validity or reliability.

1.12 Cost

Medical education assessment and selection practice is an applied science, not a pure one, therefore issues of cost are extremely significant.

1.13 Defensibility

All educational judgements are in the end arbitrary. However, they can be made defensible. The following properties are generally considered as defences against challenges: that

- Those who developed and delivered the assessment strategy were qualified to do so, both in content knowledge and assessment expertise
- Due diligence was exercised (for instance with regard to exam security)
- The methods were in accordance with accepted good practice
- The security of the exams was preserved
- Policies were applied consistently
- The policies were equitable (for instance with regard to disability)
- Clear descriptions of the process were available in advance

References

Adusumillim S, Cohan RH, Marshall KW. (2000). How well does applicant rank order predict subsequent performance during radiology residency? *Academic Radiology* **7**: 635-40.

Ainsworth and Szauter. (2006). Medical student professionalism: are we measuring the right behaviours? A comparison of professional lapses by students and physicians *Academic Medicine* **81**: S83-S86.

Albanese MA, Farrell P, Dottl S. (2005). Statistical Criteria for Setting Thresholds in medical school admissions. *Advances in Health Sciences Education* **10**: 89-103.

Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. (2003) Assessing personal qualities in medical school admissions. *Academic Medicine* **78**: 313-321.

Altmaier EM, Smith WL, O'Halloran CM, Franken EA Jr. (1992). The predictive utility of behaviour based interviewing compared with traditional interviewing in the selection of radiology residents. *Investigitive Radiology* **27**: 385-9.

Amsellem-Ouazana D, Van Pee D, Godlin V. (2006). Use of portfolios as a learning and assessment tool ia a sugical practical session of urology during undergraduate medical training. *Medical Teacher* **28**: 356-9.

Arno M M Muijtjens 1 , Lambert W T Schuwirth 1 , Janke Cohen-Schotanus 2 , Arnold J N M Thoben 3 & Cees P M van der Vleuten (2007). Benchmarking by cross-institutional comparison of student achievement in a progress test. *Medical Education* **42**: 82-88.

Bajaj G, Carmichael KD. (2004). What attributes are necessary to be selected for an orthopaedic surgery residency position: perceptions of faculty and residents. *Southern Medical Journal* **97**: 1179-85.

Bandiera G, Regehr G. (2004). Reliability of a structured interview scoring instrument for a Canadian postgraduate emergency medicine training program. *Academic Emergency Medicine* **11**: 27-32.

Barber SG, Rodd J. (1992). Can selection of medical staff be improved? *British Journal of Clinical Practice* **46**: 123-6.

Barrick MR, Mount MK (1991). The Big Five personality dimensions and job performance: a meta-analysis. *Personnel Psychology* **44**: 1-26.

Barrick MR, Mount MK, Strauss JP (1993). Conscientiousness and performance of sales representatives: Test of the mediating effect of goal setting. *Journal of Applied Psychology* **78**: 715-722.

Basco WT, Gilbert GE, Chessman AW, Blue AV. (2000). The ability of a medical school admission process to predict clinical performance and patient satisfaction. *Academic Medicine* **75**: 743-7.

Basco WT, Gilbert GE, Chessman AW, Blue AV. (2000). The ability of a medical school admission process to predict clinical performance and patient satisfaction. *Academic Medicine* **75**: 734-7.

Basco WT, Lancaster CJ, Gilbert GE, Carey ME, Blue AV. (2008). Medical school application interview score has limited predictive validity for performance on a fourth year clinical practice examination. *British Journal of Clinical Practice* **13**: 151-62.

Beley S, Dubosq F, Simon P, Larre S, Battisti S, Ballereau C, Boublil V, Richard F, Roupret M. (2005). Improvement of the recruitment of surgery interns derived from the Epreuves Nationales Classantes(National Ranking Exam): practical solution applied to urolog. *Progres en Urologie* **15**: 1101-5.

Bennett M, Wakeford R. (1983). *Selecting Students for Training in Health Care. A practical guide to improving selection procedures.* Geneva Switzerland: World Health Organisation.

Bindal T, Wall D, Goodyear HM. (2007). Performance of Paediatric Senior House Officers following changes in recruitment. *Medical Teacher* **29**: 501-3.

Blazey ME, MacLeod JA. (1996). Competency: a basis for the selection of staff nurses. *Health Care Supervision* 1**4**: 47-56.

Bloom BS (ed.) (1956) *Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain* New York: McKay

Bobek JR, Fleck MA, Moyer K, Paules CA. (1984). Selection of medical technology students in a hospital-based program. *Journal of Allied Health* **13**: 197-204.

Bobko P, Roth PL, Potosky D (1999). Derivation and implications of meta-analytic matrix incorporating cognitive ability, alternative predictors and job performance. *Personnel Psychology* **52**: 561-89.

Boor M, Wartman SA, Reuben DB. (1983). Relationship of physical appearance and professional demeanor to interview evaluations and ranking of medical residency applicants. *Journal of Psychology* **113**: 61-5.

Bore MR, Munro D, Kerridge I & Powis DA (2005). Selection of medical students according to their moral orientation. *Medical Education* **39**: 266-275.

Boursicot KAM, Roberts T, Pell G. (2007) Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education.* **41**: 1024.

Boursicot, K. A. M., 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 Science Education* **11**: 173-183.

Burgess MM, Calkins V, Richards JM. (1983). The structured interview; a selection device. Psychological Reports , 31(3):867-77.Cohen BJ. (1983). Training in general practice. How to interview candidates. *British Medical Journal* **11**: 1867-8.

Campion MA, Pursell ED, Brown BK (1988). Structured interviewing: raising the psychometric properties of the employment interview. *Personnel Psychology* **41**: 25-42.

Case S. (1992). Validity of NBME Parts I and II for the Selection of Residents: the case of orthopaedic surgery.

Chotanus J, Van Rossum H J M, Van der Vleuten C P M (1999). An Inter- and Intra-University Comparison With Short Case-Based Teaching. *Advances in Health Sciences Education* **4**: 233-244.

Clay AS, Petrusa E, Harker M, Andolsek K. (2007). Development of a web based speciality specific portfolio. *Medical Teacher* **29**: 311-6.

Cofer JB, BidermanMD, Lewis PL, Potts JR, Laws HL, Oleary JP, Richardson JD. (2001). Is the quality of surgical residency applicants deteriorating? *American Journal of Surgery* **181**: 44-9.

Cohen J. (1988) *Statistical power analysis for the behavioural sciences*. 2 ed. Lawrence Earlbaum Associates, Hillsdale, NJ.

Colbert CY, Ownby AR, Butler PM. (2008). A review of portfolio use in residency programs and considerations before implementation. *Teach Learn Med* **20**: 340-5.
Collins JP, White GR, Petrie KJ, Willoughby EW. (1995). A structured panel interview and group exercise in the selection of medical students. *Medical Education* **29**: 332-6.

Cortina JM, Goldstein NB, Payne SC, Davison HK, Gilliland SW (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology* **53**: 325-51.

Courneya A, Wright K, Frinton V, Mak E, Schulzer M, Pachev G. (2005). Medical student selection: choice of a semi-structured panel interview or an unstructured ono-on-one interview. *Medical Teacher* **27**: 499-503.

Crane JT, Ferraro CM. (2000). Selection criteria for emergency medicine residency applicants. *Academic Emergency Medicine* **7**: 54-60.

Crocker L. (1978). The role of the interview in student selection. *American Journal of Medical Technology* **44**: 438-4.

Crocker L. (1978). The role of the interview in student selection. *American Journal of Medical Technology* **44**: 438-4.

Daly KA, Levine SC, Adams GL. (2006). Predictors for resident success in otolaryngology. *Journal of American College of Surgery* **202**: 649-54.

Dannefer EF, Henson LC. (2007). The portfolio approach to competency based assessment at the Cleveland Clinic Lerner College of Medicine. *Academic Medicine* **82**: 493-502.

Davies H, Khera N, Stroobant J. (2005). Portfolios, appraisal, revalidation and all that: a user's guide for consultants. *Archives Dis Child* **90**: 165-70.

Davis MH, Friedman Ben David M, Harden RM, Howie P, Ker J, McGhee C, Pippard MJ, Snadden D . (2009). Student perceptions of a portfolio assessment process. *Medical Education* **43**: 89-98.

Davis MH, Friedman Ben-David M, Harden RM, Howie P, Ker J, McGhee C, Pippard MJ, Snadden D. (2001) Portfolio assessment in medical students final examinations. *Medical Teacher*; **23**: 357-366

Davis MH, Ponnamperuma GG, Ker JS. (2001). Portfolio assessment in medical students' final examinations. *Medical Teacher* **23**: 357-366.

Deitte L. (2008). Learning portfolios in radiology residency education: how do I get started? *Journal of American Coll Radiology* **5**: 644-9.

Dornan T, Carroll C, Parboosingh J. (2002) An electronic learning portfolio for reflective continuing professional development. *Medical Education* **36**:767–9

Downing SM, Maatsch JL. (1978). *The Effects of Clinically Relevant Multiple-Choice items on the statistical discrimination of physician clinical competence,* Michigan State University, Michigan.

Driessen E, van Tartwijk, van der Vleuten C, Wass. (2007). Portfolios in medical education: why do they meet with mixed success? A systematic review. *Medical Education* **41**: 1224-33.

Driessen EW, Overeem K , van der Vleuten,CP, Muijtjens AM. (2006). Validity of portfolio assessment: which qualities determine ratings? *Medical Education* **40**: 862-6.

Edwards JC, Johnson E, Molidor JB . (1990). The interview in the admission process. *Academic Medicine* **65**: 167-77.

Edwards JC, Johnson EK, Molidor JB. (1990) The interview in the admission process. *Academic Medicine* **65**: 167-77.

Elliot SL, Epstein J. (2005). Selecting the future doctors: the role of graduate medical programmes. *Internal Medicine Journal* **35**: 167-77.

Elwyn G, Carlisle S, Hocking P, Smail S. (2001). Practice and professional development plans (PPDs):results of a feasibility study. *Family Practice* March 27 **2**: 1.

Erlandson EE, Calhoun JG, Barrack FM, Hull AL, Youmans LC, Davis WK, Bartlett RH. (1982). Resident selection: applicant selection criteria compared with performance. *Surgery* **92**: 270-5.

Eva KW, Rosenfeld J, Reiter HI, Norman GR. (2004a.) An admissions OSCE: the multiple mini-interview. *Medical Education* **38**: 314– 26.

Eva KW, Reiter HI, Rosenfeld J, Norman GR. (2004b). The ability of the multiple mini-interview to predict pre-clerkship performance in medical school. *Academic Medicine* 2004 **79**: S40– 2.

Eva KW, Reiter HI, Rosenfeld J, Norman GR (2004c). The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. *Academic Medicine* **79**: 602– 9.

Ewan C, Melville P. (1982). Interviews of medical school entrants: to what purpose? *The Medical Journal of Australia* **2**: 233-6.

Hrisos S, Illing JC, Burford BC (2008) Portfolio learning for foundation doctors: early feedback on its use in the clinical workplace. *Medical Education* **42**: 214-223.

Ferguson E, James D, Maddeley L. (2002) Factors associated with success in medical school and in a medical career. *BMJ* **324**: 952-957

Ferguson E, James D, O'Hehir F, Sanders A (2003). Pilot study of the roles of personality, references and personal statements in relation to performance over the five years of a medical degree. *British Medical Journal* **326**: 429-431.

Ferguson E, Sanders A, O'Hehir F, James D. (2000) Predictive validity of personal statement and the role of the five factor model of personality in relation to medical training. J *Occ Org Psychol* **73**: 32144.

Fielder IG, Klingbeil G. (1991). A recruitment model for selecting residents. *Academic Medicine* **66**: 476-8.

Fish DE, Radfar-Baublitz L, Choi H, Felsenthal G. (2003). Correlation of standardized testing results with success on the 2001 American Board of Physical Medicine and Rehabilitation Part 1 Board Certificate Examination. *American Journal of Physical Medicine and Rehabilitation* **82**: 686-91.

Fish DE, Radfar-Baublitz L, Choi H, Felsenthal G. (2003). Correlation of standardized testing results with success on the 2001 American Board of Physical Medicine and Rehabilitation Part 1 Board Certificate Examination. *American Journal of Physical Medicine and Rehabilitation* **82**: 686-91.

Forker JE, McDonald ME. (1996). Methodologic trends in the healthcare professions: portfolio assessment. *Nurse Education* **21**: 9-10.

Friedman Ben David M, Davis MH, Harden RM, Howie PW, Ker J, Pippard MJ. (2001). AMEE Medical Education Guide No 24: Portfolios as a method of student assessment. *Medical Teacher* **23**: 535-551.

Friedman CP, Helm KP, Trier WC, Croom RD, Davis WA. (1991). Predictive validity of a house-officer selection process at one medical school. *Academic Medicine* **66**: 471-3.

Gadbury-Amyot CC, Kim J, Palm RL, Mills GE, Noble E, Overman PR. (2003). Validity and reliability of portfolio assessment of competency in a baccalaureate dental hygiene program. *Journal of Dental Education* **67**: 991-1002.

Gallagher AG, Neary P, Gillen P, Lane B, Whelan A, Tanner WA, Traynor O. (2008). Novel method for assessment and selection of trainees for higher surgical training in general surgery. *ANZ Journal of Surgery* **78**: 282-90.

Garden FH, Smith BS. (1989). criteria for selection of physical medicine and rehabilitation residents. A survey of current practices and suggested changes. *American Journal of Physical Medicine and Rehabilitation* **68**: 123-7.

Garner J, O’Sullivan H (2009) Medical students – in healthy competition? HEA Medev Newsletter 01.18 Spring 2009, p 5.

Gayed NM. (1991). Residency directors' assessments of which selection criteria predict the performances of foreign-born foreign medical graduates during internal medicine residencies. *Academic Medicine* **66**: 699-701.

Gayed NM. (1991). Residency Directors' Assessments of which selection criteria best predict the performance of foreign-born medical graduates during internal medicine residencies. *Academic Medicine* **66**: 699-701.

Genovich-Richards J, Berg R, Carboneau C, Molter-Sundock B, Frank G. (2005). Beyond interviewing: a21st-century recruitment process for healthcare quality professionals. *Journal for Healthcare Quality* **27**: 15-21,44.

Genovich-Richards J, Berg R, Carboneau C, Molter-Sundock B, Frank G. (2005). Beyond interviewing: a21st-century recruitment process for healthcare quality professionals. *Journal for Healthcare Quality* **27**: 15-21,44.

George JM, Young D, Metz EN. (1989). Evaluating selected internship candidates and their subsequent performance. *Academic Medicine* **64**: 480-2.

Glick SM. (1994). Selection of medical students: the Beer-Sheva perspective. *Medical Education* **28**: 265-70.

GCM (2007) http://www.gmc-uk.org/

Goldacre MJ, Turner G, Lambert TW (2004). Variation by medical school in career choices of UK graduates of 1999 and 2000. *Medical Education*. **38**: 249-58.

Gong H Jr, Parker NH, Apgar FA, Shank C. (1984). Influence of the interview on ranking in the residency selection process. *Medical Education* **18**: 366-9.

Gordon J (2003) Assessing students' personal and professional development using portfolios and interviews. Medical Education 37: 335-340.

Goodyear HM, Jyothish D, Diwakar V, Wall D. (2007). Reliability of a regional junior doctor recruitment process. *Medical Teacher* **29**: 504-6.

Gottlieb ES, Berenson M, Aries N. (1998). Optimal internship/residency interview assignments for healthcareadministration student places. *Journal of Health Administration and Education* 1**6**: 13-32.

Green HL, Beattie HM, Russo AR, Johnson A, Stickley W, Goldberg R. (1985). Impact of a residency program information videotape on resident interviewing as a time-saving strategy. *Journal of Medical Education* **60**: 335-7.

Green HL. (2009). Selection criteria for residency: results of a national program directors survey. *Academic Medicine* **84**: 362-7.

Greenwald RA, Wiener S. (1976). A standardized interviewing technique for evaluating postgraduate training applicants. *Journal of Medical Education* **51**: 912-8.

Grieshaber W. (2005). The pre-employment interview. *Radiology Management* **27**: 14-6, 18-20.

Grober ED, Matsumoto Ed, Jewett MA, Chin JL, Canadian Urology Program Directors. (2003). The Canadian Urology Fair:a model for minimising the financial and academic costs of the residency selection process. *Canadian Journal of Surgery* **46**: 458-62.

Gurza-Dully P, Melaney M. (1992). Application form items as predictors of performance and longevity among respiratory therapists: a multiple regression analysis. *Respiratory Care* **37**:137-43.

Hamel P, Boisjoly H, Corriveau C, Fallaha N, Lahoud S, Luneau K, Olivers S, Rouleau J, Toffoli D . (2007). Using the CanMEDS roles when interviewing for an opthalmology residency program. *Canadian Journal of Opthalmology* **42**: 299-304.

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

Hess TG, Brown DR. (1977) Actuarial prediction of performance in a six year A.B.-M.D. program. *J Medical Education* **52**: 68-9

Hobfollnt SE, Anson O, Antonovsky A. (1982). Personality factors as predictors of medical students performance. *Medical Education* **16**: 251-8.

Hobfollnt SE, Anson O, Antonovsky A. (1982). Personality factors as predictors of medical students performance. *Medical Education* , 16(5):251-8.

Hofmeister M, Lockyer J, Crutcher R. (2008). The acceptability of the multiple mini interview for resident selection. *Family Medicine* **40**: 734-40.

Hojat M, Mangione S, Nasca TJ, Cohen MJM, Gonnella JS, Erdmann JB, Veloski JJ, Magee M. (2001) The Jefferson Scale of Physician Empathy: development and preliminary psychometric data. *Educ Psychol Measurement* 2001 **61**: 349–65.

Holmboe ES, Wang Y, Meehan T, et al. (2008) Association between maintenance of certification examination scores and quality of care for medicare beneficiaries. *Arch Intern Med* **168**: 1396-403.

Hough LM, Eaton NK, Dunnette MD, Kamp JP, McCloy RA. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology* **75**: 581-595.

Humphreys S, Dowson S, Wall D, Diwakar V, Goodyear HM. (2008). Multiple mini interviews: opinions of candidates and interviewers. *Medical Education* **42**: 207-13.

Jackson GC. (1958). An experiment with the group interview in the selection of medical students. *Journal of Medical Education* **33**: 491-500.

Jackson GG, Kellow WF. (1958). A experiment with the group interview in the selection of medical students. *Journal of Medical Education* **33**: 491-500.

Jha V, Bekker HL, Duffy SRG, Roberts TE. (2007). A systematic review of studies assessing and facilitating attitudes towards professionalism in medicine. *Medical Education* **41**: 822–829.

Johnson EK, Edwards JC. (1991). Current practices in admission interviews at U.S. medical schools. *Academic Medicine* **66**: 408-12.

Jones A, McArdle PJ, O'Neill PA. (2002). Perceptions of how well graduates are prepared for the role of pre-registration house officer: a comparison of outcomes from a traditional and an integrated PBL curriculum. *Medical Education* **36**: 16–25.

Kalet AL, Sanger J, Chase J, Keller A, Schwartz MD, Fishman ML, Garfall AL, Kitay A. (2007). promoting professionalism through an online professional development portfolio: successes, joys and frustrations. *Academic Medicine* **82**: 1065-72.

Kalet AL, Sanger J, Chase J, Keller A, Schwartz MD, Fishman ML, Garfall AL, Kitay A. (2007). promoting professionalism through an online professional development portfolio: successes, joys and frustrations. *Academic Medicine* **82**: 1065-72.

Kesler RW, Nowacek G, Lohr JA. (1980). The high cost of recruiting residents. *Southern Medical Journal* **73**: 1521-3.

Khongphatthanayothin A, Chongsrisawat V, Wananukul S, Sanpavat S. (2002). Resident recruitment: what are good predictors of performance during pediatric training? *Journal of Medical Association Thailand* **85**: 302-11.

Komives E. (1984). The applicant interview as a predictor of resident performance. *Journal of Medical Education* **59**: 425-26.

Korman M, Stabblefield RL. (1971) Medical school evaluation and internship performance. *J Medical Education* **64**: 670-673.

Lamsdale C, Wood R, Mulrooney C. (1999) An alternative to an assessment centre on six pieces of paper? *Int J Select Assess* 1999 **7**: 170–6.

Lee AG, Golnik KC, Oetting TA, Beaver HA, Boldt HC, Olson R, Greenlee E, Abramoff MD, Johnson AT, Carter K. (2008). Re-engineering the resident application process in opthalmology: a literature review and recommendations for improvement. *Survey of Opthalmology* **53**: 164-76.

Lewis KO, Baker RC. (2007). The development of an electronic educational portfolio: an outline for medical education professionals. *Teach Learn Med* **19**: 139-47.

Longmead HE. (2003). Resident recruitment. *Academic Radiology* **10**: S4-9.

Lumsden MA, Bore M, Millar K, Jack R, Powis D (2005). Assessment of personal qualities in relation to admission to medical school. *Medical Education* **39**: 258-265.

Lynch DC, Surdyk PM, Eiser AR. (2004). Assessing professionalism: a review of the literature. *Medical Teacher* **26**: 366–373.

Mallott D. (2006). Interview, Dean's letter and Affective Domain issues. *Clinical Orthopaedics and Related Research* **449**: 56-61.

McCrorie P, Boursicot K (2009). Variations in Medical School Graduating Examinations in the United Kingdom: are clinical competence standards comparable? In press in *Medical Teacher*.

McCrorie P, Boursicot KAM, Southgate LJ. (2008) Order in variety we see; though all things differ, all agree: clinical assessment for graduation—is there equivalence across UK medical schools? *Ozzawa Conference Abstracts*: 244.

McDaniel MA, Nguyen NT. (2001) Situational judgement tests: a review of practice and constructs assessed. *Int J Select Assess* **9**: 103–13.

McDaniel MA, Whetzel DL, Schmidt FL, Maurer S (1994). The validity of employment interviews: a comprehensive review and meta-analysis. *Journal of Applied Psychology* **79**: 599-615.

McLachlan JC, Finn G, Macnaughton J (2009). "The Conscientiousness Index: a novel tool for exploring students' professionalism. In press in *Academic Medicine (May 2009).*

McLachlan JC, Finn G, Macnaughton J (2009) "The Conscientiousness Index: a novel tool for exploring students' professionalism. In press in *Academic Medicine (May 2009).*

McLachlan, JC & Whiten, S.C. (2000). Marks, scores and grades: scaling and aggregating student assessment outcomes. *Medical Education* **34**: 788-797.

McManus IC, Richards B. (1986) Prospective study of performance of medical students during pre-clinical years. *BMJ* **293**: 124-7.

McManus IC, Richards P. (1984). Audit of admission to medical school: II - shortlisting and interviews. *British Medical Journal* **10**: 1288-90.

McManus IC, Richards P. (1984). Audit of admission to medical school: III-applicants' perceptions and proposals for change. *British Medical Journal* **17**: 1365-7.

McManus JE, Elder AT, de Champlain A, Dacre JE, Mollon J, Chis L. (2008). Graduates of different UK medical schools show substantial differences in performance on MRCP (UK) Part 1, Part 2 and PACES examinations. *BMC Medicine* **14**: 6:5.

McMullan M, Endacott R, Gray MA, Jasper M, Miller CM, Scholes J, Webb C. (2003).Portfolios and assessment of competence: a review of the literature. *Journal of advanced Nursing* **41**: 283-94.

Melnick DE (2006). From defending the walls to improving global medical education: Fifty years of collaboration between the ECFMG and the NBME. *Academic Medicine* **81**: S30-S35.

Melnick DE, Dillon GF, Swanson DB (2002). Medical licensing examinations in the United States. *Journal of Dental Education* **66**: 595-599.

Melville C, Rees M, Brookfield D, Anderson J. (2004). Portfolios for assessment of peadiatric specialist registrars. *Medical Education* **38**: 1117-25.

Mensh IN. (1978). World Health Organization, Geneva (Switzerland. *Medical Education* **53**: 741-5.

Meridith KE, Dunlap MR, Baker HH. (1982) Subjective and objective admissions factors as predictors of clinical clerkship performance. *J Medical Education*. **57**: 743-51.

Mitchell G, Mitchell D, McGregor M. (1987). Selection of medical students - are interview evaluations consistent?South Af. *South African Medical Journal* **20**: 774-6.

Mitchell JR. (1978). The role of the interview in student selection. *South African Medical Journal* **54**: 932-3.

Mount MK, Barrick MR. (1995). The Big Five personality dimensions: Implications for research and practice in human resources management. In Ferris G (ed) *Research in personnel and human resources management*, JAI Press, Greenwich, CT. pp153-200.

Munro D, Bore MR & Powis DA (2005). Personality factors in professional ethical behaviour: studies of empathy and narcissism. *Australian Journal of Psychology* **57**: 49-60.

Murden R, Galloway GM, Reid JC, Colwill JM. (1978) Academic and personal predictors of clinical success in medical school. *J Medical Education*. **53**: 711-9.

Niebuhr BR. (1977). *A Structured Interview for the Selection of Physician's Assistant,* University of Texas, Galveston.

Nooman Z, Schmidt HG, Ezzat E (eds).(1990) *Innovation in Medical Education, an Evaluation of its Present Status*, Springer Publishing, New York. pp 40–9.

Nowacek GA, Bailey BA, Sturgill BC. (1996). Influence of the interview on the evaluation of applicants to interview. *Academic medicine* **71**: 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 and Learning in Medicine* **18**: 310-5.

Omar MA. (2006). Filling the Assessment gap: using a learning portfolio in international development courses. *Journal of Health Organisation Management* **20**: 74-80.

O'Sullivan PS, Reckase MD, McClain T, Savidge MA, Clardy JA. (2004). Demonstration of portfolios to assess competency of residents. *Adv Health Science Education Theory and Practice* **9**: 309-23.

Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KM, Wollersheim HC, Grol RP. (2007). Doctor performance assessment in daily practise: does it help doctors or not? A systematic review. *Medical Education* **41**: 1039-49.

Papadakis MA et al (2004). Unprofessional behaviour in medical school is associated with subsequent disciplinary action by a state medical board. *Acad. Med* **79**: 244-9.

Papadakis M et L (2005). Disciplinary action by medical boards and prior behaviour in medical school. *N. Engl. J. Med.* **353**: 2673-2682.

Papadakis MA, Arnold GK, Blank LL, Holmboe ES, Lipner RS (2008). Performance during internal medicine residency training and subsequent disciplinary action by state licensing boards. *Annals of Internal Medicine* 2008 **148**: 869-876.

Part HM, Market RJ. (1993). Predicting the first year performances of international medical graduates in an internal medicine residency. *Academic Medicine* **68**: 856-8.

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

Patterson F Ferguson E (2007) Selection for medical education and training. ASME, Edinburgh.

Patterson F, Barron H, Carr V, Plint S, Lane P. (2009). Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Medical Education*, **43**: 50-7.

Phillips WR. (1976). The interview in Family Practice Resident selection. *Journal of Family Practice* **3**: 444-5.

Pitts J, Coles C, Thomas P. (1991). Educational portfolios in the assessment of general practice trainers: reliability od assessors. *Medical Education* **33**: 515-20.

Posthuma RA, Morgeson FP, Campion MA. (2002) Beyond employment interview validity: a comprehensive narrative review of recent research trends over time. *Personnel Psychology* **55**: 1-81.

Powis D, Bore M, Munro D & Lumsden MA (2005). Development of the Personal Qualities Assessment as a tool for selecting medical students. *Journal of Adult and Continuing Education* **11**: 3-14.

Powis D, Hamilton J, McManus IC (2007). Widening access by changing the criteria for selecting medical students. *Teaching and Teacher Education* 2007 **23**: 1235-1245.

Powis DA, Neame RL, Bristow T, Murphy LB. (1988). The objective structured interview for medical student selection. *British Medical Journal* **12**: 765-8.

Powis DA, Neame RLB, Bristow T, Murphy LB. (1988) The objective structured interview for medical student selection. *BMJ* **296**: 765-8.

Prince KJAH, van Eijs PW, Boshuizen HP, van der Vleuten CP, Scherpbier AJ. (2005) General competencies of problem-based learning (PBL) and non-PBL graduates. *Medical Education* **39**:394–401.

Prince KJAH, van Mameren H, Hylkema N, Drukker J, Scherpbier AJJA, van der Vleuten CPM. (2003). Does problem-based learning lead to deficiencies in basic science knowledge? An empirical case study on anatomy. *Medical Education* **37**:15–21.

Puryear JB, Lewis LA. (1981). Description of the interview process in selecting students for admission to U.S. medical schools. *Journal of Medical Education* **56**: 881-5.

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**: 444-8.

Rao R. (2007). The Structured Clinically Relevant Interview for Psychiatrists in Training (SCRIPT): a new standardized assessment tool for recruitment in the UK. *Academic Psychiatry* **31**(6): 443-6.

Rees CE, Sheard CE. (2004). The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: the Nottingham experience. *Medical Education* **38**: 38-44.

Reiter, Harold I, Salvatori, Penny, Rosenfeld, Jack, Trinh, Kien & Eva, Kevin W (2006). The effect of defined violations of test security on admissions outcomes using multiple mini-interviews. *Medical Education* **40**: 36-42.

Remmen R et al , (2001). Effectiveness of basic clinical skills training programmes: a cross-sectional comparison of four medical schools. *Medical Education* **35**: 121-128.

Rhoton FM, Barnes A, Flashburg M, Ronail A, Springman S.(1991). Influence of anaesthesiology residents' noncognitive skills on the occurrence of critical incidents and the residents' overall clinical performances. *Academic Medicine* **66**: 359-361.

Rhoton FM. (1989). A new method to evaluate clinical performance and critical incidents in anaesthesia: quantification of daily comments by teachers. *Medical Education* **23**: 280-289.

Rhoton FM. (1994). Professionalism and clinical excellence among anaesthesiology residents. *Academic Medicine* **69**: 313-315.

Richards P, McManus IC, Maitlis SA. (1988). Reliability of interviewing in medical student selection. *British Medical Journal* **28**: 1520-1.

Richards P, McManus IC, Maitlis SA. 1988) Reliability of interviewing in medical student selection. *BMJ* **296**: 1520-1.

RMcMullan M, Endacott R, Gray MA, Jasper M, Miller CM, Scholes J, Webb C. (2003).Portfolios and assessment of competence: a review of the literature. *Journal of advanced Nursing*, 41(3):283-94., R

Roberts C, Newble DI, O'Rourke AJ. (2002) Portfolio-based assessments in medical education: are they valid and reliable for summative purposes? Medical Education 36: 899-900.

Rosenfeld JM, Reiter HI, Trinh K, Eva KW. (2008). A cost efficiency comparison between the multi-interview and traditional admissions interviews. *Advances in Health Science Education: Theory and Practice* **13**: 43-58.

Sacks MH, Karasu S, Cooper AM, Kaplan RD. (1983). The medical student's perspective of psychiatry residency selection procedures. *American Journal of Psychiatry* **140**: 781-3.

Salvatori P. (2001). Reliability and validity of admissions tools used to select students for health professions. *Advances in Health Science Education: Theory and Practice* **6**: 159-75.

Savage PE, Hinves BL. (2000). Orthopaedic specialist registrar selection: the advantages of a candidate ranking procedure at interview. *Annals of The Royal College of Surgeons* **82**: 222-6.

Schmidt FL, Ones DS, Hunter JE. (1992). Personnel selection. *Annual Review of Psychology* **43**: 627-670.

Schmidt HG, Dauphinee WD, Patel VL. (1987). Comparing the effects of problem-based and conventional curricula in an international sample. *J Medical Education* **62**: 305–15.

Schmidt HG, Vermeulen L & van der Molen HT (2006). Long term effects of problem-based learning: a comparison of competencies acquired by graduates of a problem-based and a conventional medical school. *Medical Education* **40**: 562-567.

Schuwirth L (2007). The need for national licensing examinations. *Medical Education* **41**: 1022-1023

Schuwirth L W T, Verhoeven B H, Scherpbier A J J A, Mom E M A, Cohen-Schotanus J, Van Rossum H J M, Van der Vleuten C P M (1999). An Inter- and Intra-University Comparison With Short Case-Based Teaching Advances. *Health Sciences Education* **4**: 233-244.

Shrank WH, Reed VA, Jernstedt GC. (2004). Fostering Professionalism in Medical Education: A Call for Improved Assessment and Meaningful Incentives. *Journal General Internal Medicine* **19**: 887–892.

Smilen SW, Funai EF, Bianco AT. (2001). Residency selection: should interviewers be given applicant' board scores? *American Journal of Obstetrics and Gynecology* **184**: 508-13.

Smith PE, Dunstan FD, Wiles CM. (1991). Medical school and residency performances of students admitted with and without an admission interview. *Academic Medicine* **66**: 474-6.

Smith SR. (2006). Selecting specialist registrars by station interview. *Clinical Medicine* **6**: 279-80.

Sneed NV, Edlund B, Kerr MA. (1997). Telephone interviews: a cost effective way to select faculty. *Journal of Nurse Education* **36**: 85-7.

Sparti P. (1986). Family practice resident selection using videotaped interview. *Family Medicine* **18**: 40-1.

Stern D, Frohna and Gruppen L (2005). The prediction of professional behaviour. *Med. Educ*. **39**: 75-82.

Stern DT (2006) A framework for measuring professionalism. In Stern DT (ed) (2006) *Measuring Medical Professionalism,* Oxford University Press, Oxford. pp 3-13.

Stern DT, Friedman Ben-David M, Norcini J, Wojtczak A & Schwarz MR (2006) Setting school-level outcome standards. *Medical Education* **40**: 166-172

Stewart GL, Carson KP, Cardy RL. (1996). The joint effect of Conscientiousness and self-leadership training on employee self-directed behaviour in a severe setting. *Personnel Psychology* **49**: 143-164.

Stieger S, Reips UD. (2008). Dynamic Interview Program (DIP): automatic online interviews via the instant messenger ICQ. *Cyberpsychology and Behaviour* **11**: 201-7.

StrumbergJP, Farmer L. (2008). Educating capable doctors- a portfolio approach. Linking learning and assessment. *Medical Teacher* Dec **23**: 1-5.

Swanson WS, Harris MC, Master C, Gallagher PR, Mauro AE, Ludwig S. (2005). The impact of the interview in pediatric residency selection. *Ambulatory Pediatrics* **5**: 216-20.

Tamblyn R, Abrahamowicz M, Brailovsky P, et al. (1988) Association between licensing examination scores and resource use and quality of care in primary care practice. *JAMA* **280**: 989–96.

Tamblyn R, Abrahamowicz M, Dauphinee D, et al. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA* **298**: 993–1001.

Tamblyn R, Abrahamowicz M, Dauphinee WD, Hanley JA, Norcin J, Girard N, Grand'Maison P, Brailovsky C (2002). Association between licensure examination scores and practice in primary care. *JAMA* **288**: 3019-3026.

Taylor G. (1948). The personal interview in the selection of medical students. *Journal. Association of American Medical Colleges* **23**: 171-5.

Taylor ML, Blue AV, Mainous AG, Geesey ME, Basco WT. (2005). The relationship between the national Board of Medical Examiners' prototype of the step 2 clinical skills exam and interns' performance. *Academic Medicine* **80**: 496-501.

ten Cate O (2002). Global standards in medical education – what are the objectives? *Medical Education* **36**: 602-3.

Tett RP, Jackson RN, Rothstein M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *PERSONNEL PSYCHOLOGY* **44**: 742.

Thordarson DB, Ebramzadeh E, sangiorgio SN, Schnall SB, Patzakis MJ. (2007). Resident selection: how are we doing and why? *Clinical Orthopedics and Related Research* **459**: 255-9.

Thuret G, Brouillet E, Gain P. (2005). Logistical regression of success factors in the former internship examination for medical students. *Presse Medicale* **34**: 781-5.

Thurstone LL (1934). "The Vectors of The Mind".*Psychological Review* **41**: 1-32.
Turnwald GH, Spafford MM, Edwards JC. (2001). Veterinary school admission interviews, part 1: literature overview. *Journal of Veterinary Medical Education* **28**: 111-21.

Tutton PJ. (1993). Medical school entrants: semi-structured interview ratings, prior scholastic achievement and personality profiles. *Medical Education* **27**: 328-36.

USMLE (2008) http://www.usmle.org/General_Information/CEUP-Summary-Report-June2008.PDF

Van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM, Thoben AJNM, Cohen-Schotanus J, Van Boven CPA. (2004). Cross-institutional collaboration in assessment: a case on progress testing. *Medical Teacher* **26**: 719–25.

Van Susteren T, Suter E, Romrell LJ, Lanier L, Hatch RL. (1999) Do interviews really play an important role in the medical school selection decision? *Teach Learn Med*. **11**: 66-74.

Vancouver JB. (1989). Testing for Validity and Bias in the Use of GPA and MCAT in the selection of medical students. *Academic Medicine* **65**: 694-97.

Varela JG, Scogin FR, Vipperman RK. (1999). development and preliminary validation of a semi-structured interview for the screening of law enforcement candidates. *Behaviour Sciences and the Law* **17**: 328-36.

Verwijnen GM, van der Vleuten C, Imbos T. (1990). A comparison of an innovative medical school with traditional schools: an analysis in the cognitive domain. In Vickers MD, Reeve PE. (1993). Selecting for specialist training:1. *British Journal of Hospital Medicine* **14**: 605-8.

Vickers MD, Reeve PE. (1993). Selecting for specialist training:1. *British Journal of Hospital Medicine* **14;50**: 605-8.

Vojir CP, Bronstein RA. (1983). Applicant selection procedures: a more objective approach to the interview process. *Journal of Allied Health* **12**: 95-102.

Wagoner NE, Suriano JR, Stoner JA. (1986). Factors used by program directors to select residents. *Journal of Medical Education* **61**: 10-21.

Wakeford R, Foulkes J, McManus C, Southgate L. MRCGP. (1993) Pass rate by medical school and region of postgraduate training. *BMJ* **307**: 542-3.

Wakeford R. (2002). The use of portfolios for assessment of the competence and performance of doctors in practice. *Medical Education* **36**: 918-24.

Westwood MA, Nunn LM, Redpath C, Mills P, Crake T. (2008). applicants regard structured interviews as a fair method of selection: an audit of candidates. *Journal of the Royal Society of medicine* **101**: 252-8.

Wiesner WH, Cronshaw, SF (1988). A meta-analytic investigation of the impact of interview format and the degree of structure on the validity of the employment interview. *Journal of Occupational Psychology* **61**: 275-90.

Wilkerson L, Wimmers P, Doyle LH, Uijtdehaage S (2007). Two perspectives on the effects of a curriculum change: Student experience and the United States medical licensing examination, Step 1. *Academic Medicine* 2007 **82**: S117-S120.

Wilkinson TJ, Challis M, Hobma SO, Newble DI, Parboosingh JT, Sibbald RG, Wright N and Tanner MS (2002). Medical students’ compliance with simple administrative tasks and success in final examinations: a retrospective cohort study. *BMJ* **324**: 1554-5.

Acknowledgements

We gratefully acknowledge the help of Louise Tyler in preparing the manuscript and the references.