Medical Teacher, Vol. 28, No. 2, 2006, pp. 103–116

BEME GUIDE

BEME systematic review: Predictive values of measurements obtained in medical schools and future performance in medical practice*

HOSSAM HAMDY[1], KAMESHWAR PRASAD[2], M. BROWNELL ANDERSON[3], ALBERT SCHERPBIER[4], REED WILLIAMS[5], REIN ZWIERSTRA[6] & HELEN CUDDIHY[7]

[1]Arabian Gulf University, Bahrain; [2]All India Institute of Medical Sciences, New Delhi, India; [3]Association of American Medical Colleges, Washington DC, USA; [4]Maastricht University, Netherlands; [5]University of Southern Illinois, USA; [6]Groningen University, Netherlands; [7]Monash University, Australia

ABSTRACT Background: *Effectiveness of medical education programs is most meaningfully measured as performance of its graduates.*

Objectives: *To assess the value of measurements obtained in medical schools in predicting future performance in medical practice.*

Methods:

Search strategy: *The English literature from 1955 to 2004 was searched using MEDLINE, Embase, Cochrane's EPOC (Effective Practice and Organization of Care Group), Controlled Trial databases, ERIC, British Education Index, Psych Info, Timelit, Web of Science and hand searching of medical education journals.*

Inclusion & exclusions: *Selected studies included students assessed or followed up to internship, residency and/or practice after postgraduate training. Assessment systems and instruments studied (Predictors) were the National Board Medical Examinations (NBME) I and II, preclinical and clerkship grade-point average, Observed Standardized Clinical Examination scores and Undergraduate Dean's rankings and honors society. Outcome measures were residency supervisor ratings, NBME III, residency in-training examinations, American Specialty Board examination scores, and on-the-job practice performance.*

Data extraction: *Data were extracted by using a modification of the BEME data extraction form study objectives, design, sample variables, statistical analysis and results. All included studies are summarized in a tabular form.*

Data analysis and synthesis: *Quantitative meta-analysis and qualitative approaches were used for data analysis and synthesis including the methodological quality of the studies included.*

Results: *Of 569 studies retrieved with our search strategy, 175 full text studies were reviewed. A total of 38 studies met our inclusion criteria and 19 had sufficient data to be included in a meta-analysis of correlation coefficients. The highest correlation between predictor and outcome was NBME Part II and NBME Part III,* $r = 0.72$, *95% CI 0.30–0.49 and the lowest between NBME I and supervisor rating during residency,* $r = 0.22$, *95% CI 0.13–0.30. The approach to studying the predictive value of assessment tools varied widely between studies and no consistent approach could be identified. Overall, undergraduate grades and rankings were moderately correlated with internship and residency performance. Performance on similar instruments was more closely correlated. Studies assessing practice performance beyond postgraduate training programs were few.*

Conclusions: *There is a need for a more consistent and systematic approach to studies of the effectiveness of undergraduate assessment systems and tools and their predictive value. Although existing tools do appear to have low to moderate correlation with postgraduate training performance, little is known about their relationship to longer-term practice patterns and outcomes.*

Introduction

Prediction is one of the major roles of assessment. Measurement of *outcomes* of medical education and the predictive value of these measurements in relation to on-the-job performance, i.e. postgraduate professional training and beyond, are fundamental issues in medical education that still require further study. Studies of academic success at medical school and prediction of graduates' subsequent performance have resulted in equivocal conclusions (Pearson *et al.*, 1998).

The multi-faceted and complex nature of being a doctor, combined with the diversity and multi-dimensionality of the working environment, increases the difficulty of defining and interpreting measurable and/or observable outcomes of medical education and training programs. A recent publication on identifying priority topics for conducting systematic reviews in medical education listed as one of the first priorities, ''What are the *outcomes* we should use to evaluate medical education and to what extent do measures obtained

Correspondence: Professor Hossam Hamdy, College of Medicine & Medical Sciences, Arabian Gulf University, PO Box 22979, Manama, Bahrain. Email: meddean@agu.edu.bh

*This BEME Systematic Review, complete with figures and appendices, is published as a BEME Guide: *BEME Guide no. 5: BEME Systematic Review: Predictive values of measurements obtained in medical schools and future performance in medical practice*. Hamdy *et al.* Dundee, Association for Medical Education in Europe (2006). ISBN 1-903934-30-3 (http://www.amee.org).

ISSN 0142–159X print/ISSN 1466–187X online/06/020103–14 © 2006 Taylor & Francis
DOI: 10.1080/01421590600622723

before and in medical school predict these outcomes?'' (Wolf *et al.*, 2001).

Clinical competence, in terms of outcomes of medical education, is increasingly being measured. However, as with other concepts, there is a lack of precision and clear definition. Kane (1992) defined clinical competence as ''the degree to which an individual can use the knowledge, skills and judgment associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice.'' Substantial effort has gone into defining measures of competences in basic and higher medical education. Epstein & Hundert (2002) defined professional competence, which should be the outcome of medical education programs, as: ''The habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values and reflection in daily practice for the benefit of the individual and community being served.'' This definition captures an important feature of professional competence, which described it as a habit that will need time to be developed.

The categories recorded for assessment of clinical competence in many programs have used general terms such as 'knowledge base of basic and clinical sciences, history taking, preventive care skills and ethical/legal principles'. These categories are too general to be measured precisely and to be predictive of the candidate's future performance, which is a major function of the examination. Some of the categories such as skills assess the prerequisites of performance rather than the performance itself, which includes *processes* and *outputs* (Cox, 2000).

When looking into the predictive value of assessment measures in medical schools, it is important to consider the time of measurement of outcomes along the continuum and time line of a physician's education, training and practice. Measurement can take place during or at the end of undergraduate educational programs, immediately after graduation (internship or licensure examination), during and at the end of residency training and in practice.

Ideally, an examination at the end of an undergraduate program should predict whether a student is competent and is ready for further practice and training. Competence may *be perceived* in relation to a fixed reference point or, as a dynamic process in which measurements vary in relation to the expected level of competence and the amount of relevant experience. The measurement process should prevent false negative results, i.e. failing a student who is competent and, in particular, false positive ones, i.e. passing a student who is incompetent (van der Vleuten, 2000).

Assessment of performance of medical school graduates during their first postgraduate year (internship) provides an indicator of the quality of the undergraduate curriculum and educational process, and serves as a quality assurance measure for those involved in undergraduate and early postgraduate training (Rolfe *et al.*, 1995). Medical school grades are widely accepted measures of performance quality. It is assumed, without supportive evidence, that the grades provide a basis for predicting future performance in the workplace. The more we move away from the exiting point of the education program, the more difficult becomes the measurement; the 'noise' increases (Gonnella *et al.*, 1993). Observation of performance for purposes of student appraisal in medical schools is done with the goal of extrapolating and generalizing competence that extends beyond the tasks observed.

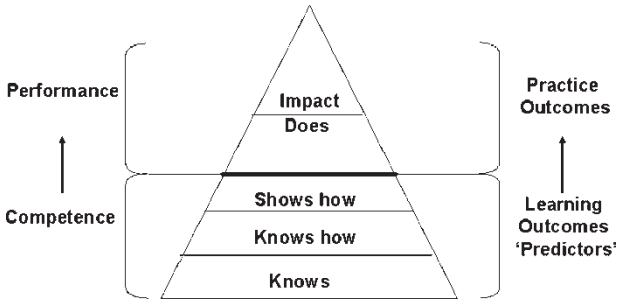


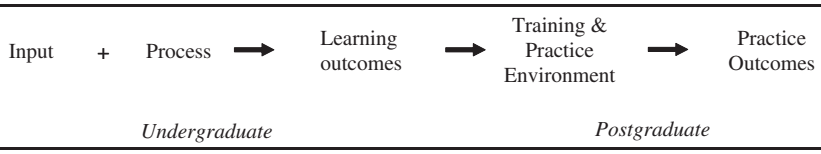
Figure 1. Conceptual relation between assessment in medical education of learning outcomes-'predictors'-and 'practice outcomes'.

Conceptual framework of the review

The prediction of performance in the real world of medical practice is now widely accepted as the goal of assessment at the different levels of medical professional education (Southgate *et al.*, 2001). A simple linear model based on measurement of input, process and output of an undergraduate medical education program cannot explain or address the complexity of measurement of its learning outcomes. The expected changes in learner behavior and performance should not only be assessed at the end of the program (learning outcomes), but more importantly in real-life practice (practice outcomes).

'Learning outcome' measurements and 'practice outcome' measurements in medical education are different. Practice outcomes of an educational program are the reflection of the program; input, process, learning outcomes and postgraduate training and practice environment. In this model, measurement of input, e.g. student characteristics, processes, e.g. educational strategy and learning outcomes during and at the end of the program, may predict to a variable degree program outcomes at different points of measurements after exiting the program, i.e. internship, residency training and on-the-job performance (practice outcomes). Time since graduation, training and practice environments have a direct impact on the physician performance and practice outcomes.

Based on the relation between 'learning outcomes' and 'practice outcomes' model, it is proposed that measurements can take place at different levels on a hierarchical pyramid based on Miller's (1990) clinical competence pyramid and Kirkpatrick's (1967) levels of effectiveness (Figure 1). It is suggested that students' 'learning outcomes' could be assessed



at three levels (knows, knows how, shows how). The assessment of medical students has tended to focus on the pyramid base 'knows' and 'knows how'. This might be appropriate in early stages of medical curriculum (Wass *et al.*, 2001), but at the end of the program higher levels should be assessed—'shows how'—which should take place at the maximum possible level of simulation to actual practice. 'Practice outcomes' need to be assessed as a performance on-the-job 'does'. The impact of the performance could be considered as the highest level of practice outcomes (Ram, 1998). This model represents a combination of assessment principles and the current development of theoretical views on medical expertise, and takes account of the difference between competence and performance (Schmidt *et al.*, 1990; Rethans 1991).

The conceptual framework that guided the systematic review is looking primarily at the predictive validity of scores or assessment of student performance in medical schools generated by different assessment systems and instruments used in measuring learning outcomes, 'Predictors', and future performance of the graduates and its ultimate impact on health, 'Practice Outcomes'.

Review question

To what extent do measurements obtained in medical schools predict outcomes in clinical practice: performance during internship, residency programs, on the job and its impact on healthcare?

Review methodology

(a) *Inclusion criteria*

For studies to be eligible for inclusion in the systematic review, they must have all of the following:

(i) Study subjects: Medical students assessed or followed up to internship, residency and/or practice after postgraduate training;
(ii) Predictors—Independent variables: 'Learning outcomes':
  (a) student ratings/scores of assessments in medical schools, preclinical and clinical phases and evaluation of student clinical competences.
(iii) Outcome variables Dependent variables: 'Practice Outcomes':
  (a) assessment scores of performance in residency or internship programs;
  (b) scores of medical licensure examination or specialty board certification examination;
  (c) health outcomes in terms of quality of life of patients, mortality or patient satisfaction, costs of care.
(iv) Study design: Studies with the following designs will be selected:
  (a) prospective follow-up study of medical students up to internship/residency/practice;
  (b) retrospective analysis of correlation between predictors and outcome variables.

(b) *Exclusion criteria*

Studies meeting the inclusion criteria will be excluded from the review if they were only reviews, contained only interim analysis of some studies with final analysis included through subsequent publication and if they were duplicate publications.

(c) *Search strategy and sources*

The search was conducted across a comprehensive range of sources in several stages. In 2001, an initial broad scoping search was performed across the key medical and educational databases, ERIC, MEDLINE, Psych Info, Web of Science and Timelit. Significant relevant papers were identified prior to this search, and strategies were drawn up to ensure each of these papers would be retrieved by the scoping search. A series of filtering strategies were developed to remove false hits (Appendix 1 on BEME website: http://www.bemecollaboration.org).

The full search included electronic and non-electronic sources. Multiple Medline searches were conducted and manageable results lists were reviewed. These searches utilized the most appropriate subject headings available, and employed limits to handle very large results sets. The Medline searches were enhanced by searches across other databases, including Embase, Cochrane's EPOC (Effective Practice and Organization of Care Group) and Controlled Trial databases, and the British Education Index.

The non-electronic search was critical in identifying papers that the databases were unable to realistically produce in manageable quantities. In addition to recommendations from experts, we also conducted hand-searches across key medical education journals: *Medical Teacher*, *Academic Medicine*, *Medical Education* and *Teaching and Learning in Medicine*.

An updating search was conducted in May 2004 to retrieve new research published since the start of the group's work. This search was limited from 2001 to the latest citations and was run across Medline, Embase, Evidence Based Medicine Reviews (including Cochrane), SPORTdiscus, AMED, HMIC, ERIC and BEI. The search strategies used were sensitive, but were not designed for maximum sensitivity, given the impracticality of the massive number of irrelevant citations that would have resulted.

The Medline search strategy combined MeSH that described aspects of the study population, predictors and outcomes; exp Professional Competence/, exp Education, Medical, Undergraduate/, Internship and Residency, Schools, Medical/, Students, Medical/Achievement/, Certification/, Educational Measurement/, Forecasting/, Longitudinal Studies/, Predictive Value of Tests/, Evaluation Studies/, Program Evaluation.

To reinforce the results of all searches, a separate cited reference search was conducted on the Web of Science. Each of the papers included for review from the first search results (as well as several from the updating search) was searched for papers that *cited it* and papers that *it cited*.

Selection methods

Selection of the studies was done by two independent reviewers applying the above criteria to papers obtained through the search strategy outlined above. Discrepancy in

the list of included studies was resolved through discussion. Inter-observer consistency was measured using kappa statistics.

Assessment of methodological quality of included studies

The methodological quality of the included studies was appraised, guided by the questions developed by BEME for assessing the quality of a research-based educational study (Harden *et al.*, 1999), using the following criteria:

(1) *Prospective or retrospective cohort*: Prospective cohort studies collect data for the purpose of correlating performance of students with their later performance as residents and practitioners. The data obtained can be complete and of high quality through the use of validated instruments and suitability for the purpose. Retrospective cohort studies have to depend on the extent and type of data collected in the past and cannot have control over their completeness or quality. Prospective studies therefore, were rated higher than the retrospective studies.

(2) *Sample*: Selection of subjects was considered unbiased if an entire batch of students is included in a prospective follow-up study or all practitioners/residents or interns or a random sample of those working in a setting are included in a retrospective study.

(3) *Data collection*:
   (a) *Similarity of correlated construct*: The degree of correlation depends on the extent of similarity between the construct. For example, clinical competence shows stronger correlation with clinical knowledge examination scores (e.g. NBME II) than with basic science examination scores (NBME I). The studies correlating clinical measures were rated higher than those correlating basic science knowledge with the clinical measures after graduation.
   (b) *Psychometric characteristics of measuring instruments*: Reliability of the instruments measuring the predictor and criterion variables affect the degree of correlation between them. Generally, the correlation tends to be alternated because the instruments are practically never perfectly reliable. Accordingly, the studies reporting the reliability of the instruments were graded higher than those not reporting it. The report, even if indicating low reliability, allows estimation of the degree of attenuation and report disattenuated correlation. Similarly, the instruments need to have established validity. Invalid instruments affect the magnitude of the observed correlation.

(4) *Data analysis*:
   (c) *Use of appropriate statistics*: The choice of correlation statistics depends on the distribution of scores and nature of relationship between the predictor and criterion variables. For example, the Pearson product–moment correlation would be appropriate if the distribution of both the predictor and criterion variables are bivariate normal and if the relationship between the two is linear.
   (d) *Attrition bias/Response rate*: Prospective studies may lose subjects during follow-up and retrospective studies are influenced by natural attrition of subjects due to summative evaluations at various stages after graduation. Thus the final set of data may represent the restricted range of eligible subjects. The degree of attrition bias/response rate may affect the magnitude of observed correlation.
   (e) *Disattenuation of correlation coefficient*: Reporting disattenuation of correlation was considered a quality issue as it would adjust the correlation coefficient for lack of perfect reliability of the scores.

Data management techniques

(a) *Data extraction*

The data extraction form was developed, pre-tested and standardized to meet the requirement of the review (Appendix 2 on BEME website: http://www.bemecollaboration.org). Two reviewers extracted the data independently. The agreement between the two reviewers was assessed using phi coefficient because Kappa gave misleadingly low values in the presence of low marginal figures. Both quantitative and qualitative data were extracted.

(b) *Data analysis and synthesis*

(1) *Criteria for addressing the combinability of the studies*
   1.1. *Similarity in timing of measurements*: Studies to be combined were selected on the basis of similarity in the timing of measurement of predictor(s) and the outcome variables.
   1.2. *Similarity in assessment methods*: The correlations between measures before and after graduation depended on the degree of similarity between the methods of assessment. For example, the knowledge in medical school assessed by objective examinations would correlate better with the knowledge assessed by objective examinations taken after graduation. Therefore, before combining the results of the independent studies, the extent of similarity of assessment methods was examined.
   1.3. *Inspection of point estimates and confidence intervals*: The quantitative data were displayed graphically as a forest plot using meta-analysis software (initially generated using Comprehensive Meta-Analysis version 1.0.23 (1998) and then transferred to SPSS for Windows version 12 to generate the actual graphs). The correlation coefficient point estimates were inspected to determine closeness to each other. The overlap of their confidence intervals was examined to determine the

extent of similarity between the results of the independent studies.

1.4. *Test of homogeneity*: Statistical tests of homogeneity 'Q statistics' were performed to examine whether the observed parameters are similar between the independent studies. A *p*-value of 0.1 or greater was accepted as indicating 'homogeneity'.

(2) *Estimating the combined correlation coefficient point estimates*

These were determined using random or fixed effects model depending on the results of the homogeneity test. If the *p*-value of the test was 0.1 or more, then a fixed effects model was used, whereas if it was less than 0.1, then a random effects model was used.

(c) *All included studies were summarized in a tabular format capturing main findings relevant to the review*

Results

Search results

Over 20,000 unique hits were returned. Their titles were visually scanned to eliminate obviously irrelevant results. In total, 560 (2.8%) titles were considered potentially relevant for abstracts review. The specificity of the scoping search (percentage of the total that was relevant to the topic) was remarkably low. While this initial search did demonstrate that the topic would prove problematic, it also revealed that there was a suitable amount of evidence to assess for systematic review.

Selection of studies

The reviewers scanned the titles and abstracts of the 569 papers retrieved by the search strategy. Of these, 175 papers were considered potentially relevant for further review. Full versions of the papers were obtained for evaluation. Two reviewers independently applied the eligibility criteria on these papers. The inter-observer agreement was substantial (kappa = 0.71; percentage agreement 89%). All papers selected for inclusion or exclusion by either of the reviewers were discussed. Differences were resolved through discussion. The screening process eliminated 137 citations that did not meet the review inclusion criteria. Thirty-eight citations were identified eligible for the review (see Appendix 3 and Appendix 6, both on BEME website: http://www.bemecollaboration.org).

Overview of the studies included in the review

Thirty-eight studies were included: one paper appeared in 1956, three in the 1960s, two in the 1970s, six in the 1980s, 15 in the 1990s and 11 in the 2000s (up to 2004).

Thirty-two studies were from the United States, three from Canada and one each from the United Kingdom, Australia and New Zealand.

Assessment of methodological quality of the studies

Two reviewers independently applied the criteria for the assessment of methodological quality of the included studies. For each of the quality criteria, the papers were rated as 'met', 'not met' or 'unclear/not applicable'. One study met all the seven validity criteria (Brailovsky *et al.*, 2001), 16 studies met four criteria and 25 met two validity criteria. Twenty-nine studies were retrospective cohort, five survey studies (Peterson *et al.*, 1956, Clute, 1963; Price, 1969; Fish *et al.*, 2003; Richards *et al.*, 1962) and four prospective cohort studies (Zu *et al.*, 1998; Probert *et al.*, 2003; Gonella *et al.*, 2004; Wilkinson & Frampton, 2004). Only one study (Brailovsky *et al.*, 2001) was prospective and reported the disattenuated correlation coefficients. Thirty-five studies had at least one construct similarity between predictor and outcome. The sample in all cohort studies consisted of an entire batch of students. All studies had a percentage of non-respondents. Only one study (Pearson *et al.*, 1998) presented a comparison of the characteristics of respondents and non-respondents. However, we analyzed all the papers, qualitatively and quantitatively, so long as they met the inclusion criteria and had relevant data.

The inter observer agreement of the quality criteria was as follows: study design = 84%, sample selection = 32%, phi = 0.15, similarity of construct = 100%, reliability of instruments = 79%, phi = 0.64, justification of statistics used = 21%, phi = −0.2, attrition/respondent bias = 21%, phi = −0.2, dissatentuation = 100%. For study design, similarity of construct and dissattentuation, phi coefficients were not estimable because one observer gave the same value to all the studies. The disagreements were due to different interpretation and unclear reporting in the studies of the quality criteria. However, all disagreements were resolved through discussion.

A wide variation was found in the methods, scales of measurements, analytical strategies and reporting styles. Psychometric characteristics of instruments were presented in a few studies. The reliability of the measures of predictor variables was given in three papers, while that for the outcome variables was given in five papers. The nature of the instruments was not described in detail. Only one paper reported reliability of both predictor and outcome variable. The most common outcome measure, 'supervisor rating', varied from study to study (e.g. scales used 25 to 33 items).

Tables in Appendix 5 (on BEME website: http://www.bemecollaboration.org) present main characteristics, results and summary of conclusions in the 38 studies included.

There was a large variation in the method of analysis as well. Nineteen studies reported the Pearson correlation coefficient and 13 had regression analysis. The correlation coefficients in the 19 studies formed the basis of our meta-analysis. They were organized around three outcome variables (Figures 2–11 on BEME website: http://www.bemecollaboration.org).

(a) *Supervisor rating during residency*: was the outcome variable in eight studies, four had NBME I and NBME II as predictors, two had clerkship GPA as predictors and two had all the three predictors. Results for each

predictor are presented below:

(i) NBME I: The Pearson correlation coefficients of the nine data sets in eight studies are plotted in Figure 2 (on BEME website: http://www.bemecollaboration.org). All the point estimates indicate a positive correlation. The confidence intervals of the results overlap with each other. The test of heterogeneity is significant (Q value = 25.166, df = 8, $p = 0.0015$) indicating lack of homogeneity of the results, but visual inspection of the point estimates and the confidence interval indicates an acceptable level of similarity in the results. The combined results (using random effects model) yielded low correlation (Pearson $r = 0.22$; 95%, CI 0.13–0.30).

(ii) NBME II: Results from seven data sets in six studies are shown in Figure 3 (on BEME website: http://www.bemecollaboration.org). All the point estimates indicate a positive correlation. The confidence intervals of the five studies (except Smith, 1993) overlap with each other. The test of heterogeneity is also significant (Q value = 26.539; df = 6, $p = 0.0002$). The summary correlation coefficient using random effects model is 0.27 (95%, CI 0.16 to 0.38) indicating a low correlation.

(iii) Clerkship GPA: The results from 11 data sets in 10 studies are shown in Figure 4 (on BEME website: http://www.bemecollaboration.org). All the point estimates indicate a positive correlation in all the studies with overlapping confidence intervals. The test of heterogeneity is also significant (Q value = 46.87, df = 10, $p = 0.0005$) but visual inspection of the point estimates and confidence intervals indicates an acceptable level of similarity in the results across the studies. The combined correlation coefficient using random effects model showed a low correlation (Pearson $r = 0.28$, 95%, CI 0.22–0.35).

(iv) OSCE: Five studies had OSCE as the predictor variable as shown in Figure 5 (on BEME website: http://www.bemecollaboration.org). The correlation coefficients were similar. Test of heterogeneity was non-significant (Q value = 1.0267, df = 3, $p = 0.7948$). The combined correlation coefficient using fixed effects model was low (Pearson $r = 0.37$; 95%, CI 0.22–0.50).

(v) Ranks based on Dean's letter: Three data sets from two studies lent themselves to meta-analysis for this predictor. The correlation coefficients were similar as shown in Figure 6 (on BEME website: http://www.bemecollaboration.org). The test of heterogeneity was non-significant (Q value = 0.024, df = 2, $p = 0.988$). The combined estimate of correlation coefficient using fixed effects model indicated low correlation (Pearson $r = 0.22$; 95%, CI 0.12–0.31).

(vi) Preclinical GPA: Only four studies, five data sets had this predictor with supervisor rating as the outcome. All the point estimates indicated positive correlation of similar magnitude as shown in Figure 7 (on BEME website: http://www.bemecollaboration.org). The confidence intervals were overlapping. The test of heterogeneity was non-significant (Q value = 0.7399, df = 4, $p = 0.9463$). The combined estimate, using fixed effects model, indicated low correlation (Pearson $r = 0.25$; 95%, CI 0.19–0.31).

The reliability of the measuring scale of the supervisor ratings were given in four studies: Markert (1993) 0.964; Paolo *et al.* (2003) 0.98; Fine & Hayward (1995) 0.8, Hojat *et al.* (1986) 0.86.

(b) *NBME III*: Two studies correlated NBME I and II with NBME III scores. Both had a large sample size: 628 for Markert (1993) and 2368 for Hojat *et al.* (1993). Between NBME I and NBME III, the correlation coefficients, in the two studies, were similar as shown in Figure 8 (on BEME website: http://www.bemecollaboration.org). Test of heterogeneity was statistically non-significant (Q value = 1.798, df = 1, $p = 0.18$). The combined correlation coefficient using fixed effects model was 0.59 (95% CI 0.57–0.61). Between NBME II and NBME III, the correlations were similar as shown in Figure 9 (on BEME website: http://www.bemecollaboration.org). The test of heterogeneity was statistically non-significant (Q value = 0.207, df = 1, $p = 0.649$). The combined correlation coefficient based on a fixed effects model was (Pearson $r = 0.72$; 95%, CI 0.70–0.73). These coefficients were substantially higher than those seen with the outcome supervisor rating. One study with six data sets correlated clerkship examination scores and NBME III (Rabinowitz & Hojat, 1989). The correlation coefficient between different clerkship scores (predictor) and NBME III ranged between $r = 0.32$ and $r = 0.49$.

(c) *American Board of Specialty Examinations*: Three studies, five data sets correlated NBME I scores as predictor and American Board of Specialty Examination as outcomes lent themselves for meta-analysis. The point estimates were close to each other and confidence intervals were overlapping. The test of heterogeneity was non-significant (Q value = 6.86, df = 3, $p = 0.076$). The combined correlation coefficient as shown in Figure 10 (on BEME website: http://www.bemecollaboration.org), using a fixed effects model, was moderately good (Pearson $r = 0.58$; 95%, CI 0.54–0.62).

One study (Figure 11 on BEME website: http://www.bemecollaboration.org) with three data sets correlated NBME II scores as predictor and American Board of Medical Specialty examination scores. Point estimates were close to each other and confidence intervals overlapping. Test of heterogeneity was significant. The combined correlation coefficient using random effect model was moderately good (Pearson $r = 0.61$, 95%, CI 0.51–0.70).

The studies reviewed and the meta-analysis showed that the correlations between the predictor variables of assessment in undergraduate medical education and supervisor ratings were lower than with NBME I and II as predictors and NBME III and American Board of Medical Specialty

Examinations as outcomes, although they were both statistically significant.

Some of the main findings in other studies with predictors and outcomes not included in the meta-analysis are summarized as follows.

The only study (Brailovsky *et al.*, 2001), giving disattenuated correlation coefficients showed moderate to high correlation between script concordance scores at the end of clerkship and clinical reasoning at the end of residency training.

Clerkship honor grades and honors society (Kron *et al.*, 1985; Amos & Massagli, 1996), student rank (Blacklow *et al.*, 1993; Amos & Massagli, 1996) and clerkship GPA (Arnold & Willoughby, 1993) predicted residency clinical performance and passing the written boards on the first attempt. Overall GPA in medical schools can predict performance in internship (average $r = 0.4$), (Fincher *et al.*, 1993).

The large study (6656 medical students) by Gonella *et al.* (2004) examined the predictive value of number of grades in medical schools and performance on USMLE III and supervisor rating during residency year one. They concluded that ratings of clinical competence beyond medical schools are predictive by number grades in medical schools.

The only recent study (Tamblyn *et al.*, 2002) on predicting process of care from final-year MD examination scores showed statistically significant association (Table 33, Appendix 5—on BEME website: http://www.bemecollaboration.org).

Discussion

This systematic review was conducted to determine to what extent measurements obtained in medical schools can predict outcomes in clinical practice; performance during internship, residency programs, on the job and their potential impact on health. The effectiveness of medical education programs is inseparable from the effectiveness of their products and the most meaningful measure of effectiveness is performance in practice.

Search and selection of studies

Retrieving evidence for a systematic review in medical education is problematic, and the work done by the BEME Collaboration has highlighted the difficulties of systematic searching within the educational discipline (Haig & Dozier, 2003). The sources of evidence that contain medical education research (primarily databases of peer-reviewed literature) are either medical or educational; they rarely describe medical education content adequately—and frequently even lack the descriptors to do so. In this review the search strategies were made less sensitive (to reduce the number of false hits); some of the highly relevant papers identified were invariably missed.

Additional methods were therefore required to augment the search. The group used a variety of proven methods (hand-searches, experts in the field and cited reference searches) to improve the comprehensiveness of the retrieval. Achieving a measure of absolute saturation is rarely possible when systematically searching for evidence, but there are methods to realize when further searching is likely to be redundant. One such example is reaching the point where a cited reference search no longer produces unseen results from either the initial paper or any of its derivatives. Although the review topic proved challenging, these additional methods employed ensured that the outcome was systematic and most probably comprehensive.

This BEME review group highlighted these problems. A systematic review requires a comprehensive search of all relevant sources, yet without satisfactory descriptors the searching proved difficult. Where adequate descriptors did exist (e.g. for the undergraduate population) they were applied sporadically while other key concepts had no satisfactory descriptor (e.g. predictive values/indicators of future performance). Given enough resources it would have been possible to sift through results lists numbering in the tens of thousands, but this is unrealistic for a single review group. Indeed, even with a strategy designed for maximum sensitivity it is unlikely that all relevant citations for this topic would be retrieved. However, several problems were encountered. Of the 175 potentially relevant reviewed studies, only 38 were found suitable for inclusion despite multiple strategies used to identify relevant studies. The inclusion criteria were identified in the light of the conceptual design of the study looking mainly at the relation between measurement of learning outcomes as predictors and practice outcomes including practice during residency training and beyond. This approach led to the exclusion of other predictors such as psychosocial characteristics of students and other measures of outcomes, like board certification status, medical school faculty appointments, speed of career progression, research publications and stress burnout and satisfaction. All these have been reported in the literature as indicators of a physician's professional success (Hojat *et al.*, 1997; West, 2001; McManus *et al.*, 2003).

Assessment of the methodological quality of the studies

Lack of standard reporting style of the results and their statistical analysis made it difficult to rank the studies according to their quality criteria. A quality criterion may or may not be met based on the reporting style. The reporting is particularly limited in four elements of study quality: (a) the psychometric characteristics of the measures of predictors and outcomes; (b) justification of statistics used; (c) comparison of characteristics of respondents and non-respondents; and (d) disattenuation. The question of whether a study was of poor quality or the authors did not consider it important to report some of the above elements is difficult to resolve. Recent studies have begun to address the above limitation (Brailovsky *et al.*, 2001). This review points to the need for regular reporting of the above elements of study quality in correlation studies. Strict application of a high quality threshold would have excluded a large number of the studies included. However, we have followed an inclusive approach as recommended by BEME.

Relation between predictors and outcomes

The results of the systematic review indicated that analysis of performance after graduating from medical school is complex and cannot be susceptible to one type of measurement. Clinical competence is not unidimensional. It is a multifaceted entity, and the strength of relationships with medical school performance measures varies depending on conceptual relevance of the measures taken during and after medical school (e.g. the preclinical GPAs yield more overlap with physicians' medical knowledge than with physicians' interpersonal skills). This expected pattern of relationship has been confirmed empirically (Hojat *et al.*, 1986; Hojat *et al.*, 1993).

Medical education changes across the continuum of undergraduate medical education, postgraduate training and unsupervised practice performance. The latter is difficult to measure and predicting practice performance years after formal training is even more difficult. In the 1960s studies showed a lack of relationship between medical school academic performances (GPA) and practice performance (Price, 1969). Although lack of correlation was explained on the basis of restriction of range, Price *et al.* (1964) argued that it was relatively unimportant.

In the 1970s, Price *et al.* (1973) and Wingard & Williamson (1973) published two evaluative literature reviews on grades as predictors of physicians' career performance. The main findings of these reviews indicated that, at that time, very little data on this subject existed with little or no correlation between the two factors. However, these studies were limited by the type of measures used to predict performance in practice.

In the 1990s Taylor & Albo (1993) studied physicians' performance and their relationship to two predictors: performance of medical students in their academic years (one and two) and their clinical years (three and four). In this study correlation between 167 physicians' medical school grades and 61 composite performance scores ranged from –0.25 to 0.28. This poor correlation also applied to sub-groups based on the number of practice years and specialties. This position—that performance during medical school does not differentiate applicants who will perform well during residency from those who will perform poorly—was supported by Brown *et al.* (1993) and Borowitz *et al.* (2000). These studies indicated that the complex competences needed for a physician to perform effectively are poorly measured by academic scores obtained through measurements that examine a narrow band of the extremely complex total spectrum of skills, abilities and performances of practicing physicians.

There have been many explanations for the weak association between medical school and postgraduate performance and the inconsistent findings of previous research (Wingard & Williamson, 1973; Gonnella *et al.*, 1993). These include deficiencies in traditional grading systems or an inherent inability of grades to indicate the transformation of potential into the workplace, the effect of intervening experience between the time of academic training and subsequent career evaluation, and the failure of the selection processes of traditional medical schools to identify students with the characteristics that might be prerequisite for successful performance (changing mindsets: knowledge, skills, behaviors, and professionalism) in the work environment (Pearson *et al.*, 1998).

The correlation between performance measures in medical school and in practice is always an under-estimated index of relationship, because of the exclusion of those in the lower tail of performance distribution in medical school due to attrition. Attrition always restricts the range of grade distribution, leading to less overlap and shrinkage of correlations. This and other conceptual and methodological issues involved in predicting physician performance from measures of attainment in medical school have been reported (Gonella *et al.*, 1993).

Other researchers have established a moderate relationship between academic performance at the medical school and practice performance, with higher correlations when an attribute is evaluated by a similar assessment method (Hojat *et al.*, 1993; Markert, 1993).

Predicting performance during postgraduate training

In this systematic review we were able to combine in a meta-analysis the correlation coefficients from only 19 of the included 38 studies. This was due to:

(1) Variability of the measured predictors in medical schools: 25 variables could be identified from the studies included. Some had objective measurements, e.g. NBME/USMLE scores, and other subjective measurements, e.g. ranking using Dean's letter or Honours Society 'AOA'.
(2) Variability of the outcomes and how they were measured. Four outcome measures were identified in the studies included in the meta-analysis, NBME III, supervisor's ratings during internship and different years of residency training, in-training examination of residents, and American Board of Medical Specialties Examination.

The meta-analysis demonstrated that summary correlations between NBME/USMLE I and supervisor rating during internship or first year residency was low (0.22), though statistically significant and consistent with the previous longitudinal study data of Hojat *et al.* (1993), and Gonnella *et al.* (1993). However, correlation of NBME I and NBME II with NBME III and American Board of Specialty Examinations was moderately high (0.6–0.7) and statistically significant.

Although significant improvement is taking place in student assessment in the clinical years, the problem of measurement of clinical competence of physicians in training is a complex and daunting task. The complexity of professional competence necessitates the use of multiple assessment methods to evaluate performance. Despite the availability of several evaluation tools, how objective resident supervisors are concerning the evaluation of the clinical performance of their trainees remains unclear (Holmboe & Hawkins, 1998).

It may be debatable whether specific assessment instruments such as the OSCE should be included in the systematic review. We believe that OSCE is an important instrument relatively recently incorporated in the assessment of medical students and its predictive validity should be assessed.

In this systematic review, the correlation coefficient between OSCE and supervisor rating yielded a summary estimate of 0.30 (95%, CI 0.24–0.37) suggesting a low correlation. The weak correlations, although statistically significant, obtained from several studies looking into the predictive value of constructs assessed by OSCE such as interpersonal skills, data collection and physical examination skills and residents' supervisors' rating could be explained on the basis that assessment of residents does not evaluate objectively the same constructs as assessed by the OSCE in the undergraduate program. Another explanation could be the failure to correct for disattenuation.

The study by Rutala *et al.* (1992) showed that OSCE scale was the best predictor of performance rated by the residency directors. The highest correlation was that evaluating interpersonal skills, $r = 0.42$. Other OSCE domains had a lower positive correlation, differential diagnosis $r = 0.28$, decision-making $r = 0.28$. The study by Probert *et al.* (2003) on 30 medical students demonstrated that OSCE showed consistent positive association with consultant ratings of their performance at the end of the pre-registration year. Other improved methods of assessment of clinical competences in medical schools, such as the post-clerkship clinical examination PCX, have demonstrated that the correlation with first-year residency supervisors' ratings ranged from 0.16 to 0.43, mean 0.32 (Vu *et al.*, 1992).

Recent reports from the longitudinal study of the Jefferson Medical College showed that the number grades in medical schools can predict performance in medical licensure exams and clinical competence ratings in the first postgraduate year (Gonnella *et al.*, 2004).

Some studies explored how cognitive factors (data gathering and analysis skills, knowledge, first- to fourth-year GPA and NBME I and II) and non-cognitive factors (interpersonal skills and attitudes) assessed during medical student training predicted postgraduate clinical competence (Heneman, 1983; Martin *et al.*, 1996). These studies showed that cognitive factors can account for up to 51% of the variance in NBME III grade (Markert, 1993).

Our results indicated the importance of measurements of similar constructs in order to find a positive and strong association. The correlation between clerkship GPA as predictor and supervisor rating during residency as outcome ($r = 0.3$) was higher than other predictors in the preclinical phase (NBME I $r = 0.18$). Studies in the 1960s and 1970s supported the view that grades and evaluations during clinical clerkships correlated well with performance during residency (Gough, 1963; Richard *et al.*, 1962), particularly in the clerkship related to the field of residency chosen by the student (Keck *et al.*, 1979). Another predictor that was not included in our study is evaluation by peers, which was found to be a better predictor of future internship success than were estimated by preclinical and clinical faculty (Korman & Stubblefield, 1971).

The study by Brailovsky *et al.* (2001) on script concordance between students and final-year residents, demonstrated the importance of measurements of similar constructs at two different levels of expected performance (medical students and final-year residents) along the continuum of medical education and practice. In this study, scores obtained by students at the end of clerkship using a script concordance (SC) test predicted their clinical reasoning performance at the end of residency measured by OSCE, short-answer management problems and simulated office orals. They reported generalizability coefficiencies for OSCE 0.717 ($n = 181$), short-answer management problems 0.816 ($n = 769$), and simulated office orals 0.478 ($n = 769$).

Predicting on-the-job practice performance

The complex nature of measuring performance in practice should consider the conceptual difference between competence and performance of physicians: competence 'what a doctor is capable of doing', i.e. under test conditions, and performance 'what he or she actually does in day-to-day practice' (Rethans, 1991). This concept was further described by Epstein & Hundert (2002) when defining professional competence as the habitual application of knowledge, skills and attitudes in the care of patients. Competence and performance criteria are structural and procedural measures, thus representing moderate variables in the sense of surrogates for relevant and ultimate end points of measurement: the outcome criterion 'improvement or maintenance of patient's health'.

Assessing the quality of healthcare as a reflection of the impact of physician performance is complicated. Donabedian (1998) argues that in measuring quality we need to assess not only the performance of practitioners but the contributions of patients and family, the structural attribute of the healthcare setting, the process of care and its outcomes.

In this systematic review we were able to find few studies that looked into the relationship between medical school measurements and on-the-job performance beyond residency. Four studies fulfilled our inclusion criteria (Peterson *et al.*, 1956; Clute, 1963; Price *et al.*, 1964; Tamblyn *et al.*, 2002). Tamblyn *et al.* (2002) investigated scores on Canadian licensure examinations taken immediately at the end of medical school and prediction of clinical behaviors 4–7 years later. This study was included in our systematic review as the Canadian licensure examination could be considered similar to a final-year MD examination, which measures students' learning outcomes at the point of exiting from the program. The study showed that scores on the Canadian licensure examination were a significant predictor of practice performance. In this study indicators of practice performance were selected on the basis of unexplained practice variations, and/or their association with the outcomes or costs of care: e.g. (1) mammography screening rate was used to assess preventive care; (2) continuity of care because of its importance in prevention and chronic disease management; (3) the differences between disease-specific and symptom-relief prescribing rate and contra-indicated prescribing rate; (4) contra-indicated prescribing, which accounts for 20% of drug-related adverse events; and (5) consultation rate was used as an indicator of resource use because referral determines access to higher cost specialty care.

Assessing the relationship between examination scores and more objective measures of quality of care is difficult due to the complexity of evaluation of optimal and actual practice. Setting standards of practice and its measurement should not only consider quantitative data obtained from assessment scores commonly obtained from examinations that attempt to measure 'competence' but should consider qualitative outcomes, such as patient satisfaction, efficiency, outcome of

consultation and impact of health education (Prideaux *et al.*, 2000, Tamblyn *et al.*, 1994). The process of care could also be considered as reflection of performance, e.g. screening and preventive services; diagnosis and management; prescribing; counseling and condition-specific processes of care (e.g. whether diabetics receive foot exams).

One of the main problems with studying postgraduate clinical performance is establishing a comparable scoring system for assessing competence in the different specialties. This is known as the 'criterion problem' and confronts the predictions of success in all jobs, not only medicine (Ferguson *et al.*, 2002). One solution to this problem has been to develop competence-based models of care and specific skills through detailed job analysis of individual medical specialties (Viswesvaran *et al.*, 1996; Patterson *et al.*, 2000).

Instruments used in measuring performance of residents and practicing physicians should have an acceptable degree of validity and reliability. Global rating, which forms the primary basis for appraising clinical skills, suffers from several sources of bias that involve cognitive, social and environmental factors, which affect the rating, not only the instruments. Research showed that patterns of measuring instruments account for no more than 8% of the variance in performance ratings (Williams *et al.*, 2003).

Standards of practice should be developed in relation to a core of common health problems or presentations encountered in the specific domain of practice. Sampling performance in relation to a core of health problems and health indicators should allow generalization of the results and avoid restricting the assessment to a small number of patients. Measurement of performance should not be limited to technical aspects and knowledge, but should also consider attitudes (Tamblyn, 1994). An interesting study (Papadakis *et al.*, 2004) looked into the unprofessional behavior of students in medical school and whether it is associated with subsequent disciplinary action by a state medical board. It was found that the prevalence of problematic behavior was 38% in the cases and 19% in the controls (odds ratio 2.15). These findings indicated the importance of professionalism as an essential competence to be demonstrated by a student to graduate from medical school.

Personal and psychosocial attributes are important facets of the physician's clinical competence that few empirical studies have looked into. With regard to issues of psychosocial predictors of the academic and clinical performance of medical students, they found that selected psychosocial attributes could significantly increase the validity of predicting performances on objective examinations (Hermen *et al.*, 1983; Hojat *et al.*, 1988, 1993). Hojat (1996) suggested that a significant link exists between selected psychosocial measures and physician clinical competence. Although assessing psychosocial attributes of the medical students was not part of the inclusion criteria in our systematic review, it is important to be considered and needs to be studied further.

The studies included in the systematic review provided evidence to support a relationship between measurements used in medical school and performance during residency. The magnitude of the correlation was higher when the predictors and outcomes measurements were based on objective written examination, e.g. NBME/USMLE I, II and III. On the other hand, Fine & Hayward (1995) suggested that academic performance measures have been over-emphasized as predictors of physicians' performance in residency training.

Recent developments in outcome measurements in medical education

During the late 1990s the issue of measurements of educational outcomes of undergraduate medical education and postgraduate residency training programs became an important international activity of several organizations responsible for medical education. This global activity is trying to look into three basic questions related to quality medical education: 'What to measure?'; 'How best can we measure?' and 'Is there a relation between what is measured and quality of practice?'

In the US, the Accreditation Committee of Graduate Medical Educators (ACGME), the American Board of Medical Specialties (ABMS) and the American Association of Medical Colleges (AAMC) adopted six general competences for evaluating residents and practicing physicians. The American Association of Medical Colleges (AAMC) and its accreditation committee (LCME) linked the medical school objectives to these competences, recognizing them as learning outcomes, but at a lower level of expectation than that of the residency programs (Stevens, 2000). In Canada, the Royal College of Surgeons (RCS) has developed CanMed 2000, which defines the expected competences of residency programs. In Europe the General Medical Council (GMC) in the UK and the Royal Colleges have restructured their residency programs. The World Federation for Medical Education (1998, 2003) developed global standards for basic medical education, postgraduate training and continuing professional development.

In the Middle East the committee of Deans of Medical Colleges, 'Fourteen Colleges' in six Gulf States—United Arab Emirates, Saudi Arabia, Qatar, Oman, Kuwait and Bahrain—developed accreditation standards and outcomes of undergraduate medical education (Guideline on Minimum Standards for Establishing and Accrediting Medical Schools in the Arabian Gulf Countries, 2001). The World Health Organization (WHO) Eastern Mediterranean office (EMRO) is leading a multinational project in the region to develop standards for accreditation of medical schools.

Defining global core of learning outcomes for undergraduate medical education, postgraduate residency training and continuing professional development should be organized around similar constructs. The six competences of ACGME—'Patient care, knowledge, ethics and professionalism, communication skills, practice-based learning and system-based practice'—can be a model for such constructs, which could be measured at different levels and phases of the professional life of a physician. The Dreyfus & Dreyfus (2001) taxonomy of levels of performance, which include novice, competent, proficient, expert and master, have the implication of progressive proficiency and can help in measuring performance at the end of medical school, residency training and beyond. The subjectivity of this taxonomy requires the identification of descriptors to improve its objectivity and valid, reliable instruments of measurements.

We hope that the conceptual model of this systematic review and its findings can provide the best available evidence on the predictive values of current assessment measurement in medical schools and future performance in medical practice, which should be considered in the measurement of quality in medical education.

Limitations of the systematic review

(1) An important limitation of this review is the language bias. It is highly probable that there are publications on the topic elsewhere (Liu *et al.*, 1990). Future cooperation with colleagues can help in reviewing publications in French, Spanish and German in future updating of the review, which we hope to do.

(2) The results of this systematic review were based on studies mainly from the USA and the assessment systems reported are used only in the USA, such as NBME/USMLE, honors societies (AOA) and Dean's letters. This raises the issue of generalizability of the predictive validity of the assessment measurements. On the other hand, it is possible to find similarity that could be generalized when looking at the construct to be measured, e.g. NBME I = basic medical science knowledge; NBME II = application of knowledge in clinical sciences.

(3) Meta-analysis of regression coefficients from various studies was not done because the reported regressions did not adjust for the same variables across different studies.

Future directions

(1) This systematic review emphasized the problems in retrieving evidence for medical education as a whole. The importance of employing additional methods to enhance the standard approach of searching a few core databases cannot be underestimated. While these methods will obviously require additional skills, time and resources, they are vital to ensuring that the systematic review is based on all available evidence. Not only are these additional methods more effective than trying to process massive lists of false hits but they will almost certainly return relevant results that databases currently cannot.

Although the coverage and description of medical education content has improved considerably in the last few years, there is substantial room for further improvement. By drawing attention to these challenges, and continuing to make efforts, currently under way, to overcome them (METRO Project, 2004), the BEME Collaboration can make a significant contribution to improving accessibility to the available evidence in medical education.

(2) The review identified some common measures of performance in practice beyond residency training that might be considered for future studies. These include patient outcomes and impact of the performance on health, such as mortality and morbidity of common health problems in a given community; newer outcomes like patient satisfaction, functional status of patients, cost effectiveness of management or intermediate outcomes like better control of diabetes, HbA_{1c} and lipid levels of diabetics may give indirect indication of physician performance and its impact.

(3) Similarity of the data-collection methods and statistical analysis of the results will help in increasing the homogeneity between the research results and will allow for their combinability, which will increase the strength of the evidence. It is recommended that studies should:
   (a) report reliability of the data collection method 'measurement instrument';
   (b) use similar statistical analysis, e.g. Pearson's correlation with report of confidence interval;
   (c) consider cognitive, social and environmental sources of bias in performance ratings in developing measurement instruments;
   (d) report disattenuated correlation coefficients;
   (e) whenever there are attrition and/or non-respondents in the studies, a comparison of characteristics of respondents and non-respondents need to be presented to allow assessment of attrition/respondent bias.
   (f) report the justifications of statistics used. For example, while using Pearson's correlation, an indicator of whether the relationship between predictor and outcome is linear and their distribution is bivariate normal needs to be given.

(4) Medical schools and residency training programs need to conduct longitudinal studies on their graduates. The Jefferson study is a model for this type of research (Gonnella *et al.*, 2004).

Conclusion

(1) The studies included in the review and meta-analysis provided statistically significant mild to moderate correlations between medical school assessment measurements and performance in the internship and residency. Basic science grades and clinical grades can predict residency performance.

(2) Performance on similar measurement instruments is better correlated such as:
   - NBME II scores with NBME III scores;
   - medical school clerkship grades and supervisor rating of residents;
   - OSCE and supervisor rating of residents when similar constructs are assessed.

(3) No consistent statistical analysis was used in reporting the relationship between the predictors and outcome variables. Only a few studies reported reliability of the measurement instruments and disattenuation. The methodological shortcomings of past research in testing predictive validity need to be addressed and sound models for assessing it need to be studied further, e.g, longitudinal profile development, cross-validation and inspection of the adjusted R^2 (Renger & Meadows, 1994).

(4) Evidence on predictors of performance in practice beyond residency training is rare and weak. New

measures of performance in practice, such as 'patient outcomes' and 'process of care', might be considered for future studies.

(5) The difficulty in searching encountered in this systematic review indicated the importance that medical education journals should place on encouraging the use of an agreed controlled vocabulary: keywords and MeSH terms that describe instruments and variables used in student and physician assessment and in reporting outcomes of medical education.

Contributions

H. Hamdy and K. Prasad developed the systematic review protocol and worked on all aspects of the review from conceptualization to the writing of the final manuscript. M.B. Anderson contributed to the literature search and writing of the review. A. Scherpbier, R. Williams, R. Zwierstra conducted the hand search and advised on the methodology and writing of the review. H. Cuddihy contributed to the literature search and writing of the review.

Funding

This project was funded through the Research Fund of the Arabian Gulf University, Bahrain.

Ethical approval

Approval was given by the BEME Organization and the Research and Ethics Committee of the Arabian Gulf University in Bahrain.

Acknowledgements

The authors are grateful to Mr Alex Haig, Information Scientist, Lister Postgraduate Institute, Edinburgh for his help in developing the search strategy and literature search. Special thanks go to Professor Joseph Gonella, Director, Center for Research in Medical Education and Health Care, Thomas Jefferson University, USA and Professor Gordan Guyatt, Clinical Epidemiology and Biostatistics and Medicine, Faculty of Health Sciences, McMaster University, Canada, for their valuable comments. We would also like to thank Mr Emad Maswadi for his valuable statistical support and in producing the plot graphs.

The authors would like to thank all the experts in Medical Education and the reviewers who provided them with constructive advice.

Special thanks are offered to Mrs Aldina Archibald and Mrs Olive Wagstaff for the typing of the manuscript.

Notes on contributors

HOSSAM HAMDY, MD MCh FRCS PhD, is Professor in the Department of Surgery & Medical Education, Dean, College of Medicine & Medical Sciences, Arabian Gulf University, Bahrain.

KAMESHWAR PRASAD, MD DM MSc, is Professor in the Department of Neurology, All India Institute of Medical Sciences, New Delhi, India.

M. BROWNELL ANDERSON, MEd, is Senior Associate Vice President, Medical Education, Association of American Medical Colleges, Washington DC, USA.

ALBERT SCHERPBIER, MD PhD, is Professor and Scientific Director, Medical Education Institute, Faculty of Medicine, Maastricht University, The Netherlands.

REED WILLIAMS, PhD, is Professor and Vice Chairman for Educational Affairs, Department of Surgery, University of Southern Illinois, USA.

REIN ZWIERSTRA, MD, is Director of Institute for Medical Education, Faculty of Medical Sciences, Groningen University, The Netherlands.

HELEN CUDDIHY, MD CCFP FRACGP PhD, is in the Department of Epidemiology & Preventive Medicine, Monash University, Victoria, Australia.

References

ACCREDITATION COUNCIL FOR GRADUATE MEDICAL EDUCATION (ACGME) (2000) Outcome Project: General Competencies. [Online]. Available at: http://www.acgme.org/Outcome.

ALEXANDER, G.L., DAVIS, W.K., YAN, A.C. & FANTONE, J.C. (2000) Following medical school graduates into practice: residency directors' assessments after the first year of residency, *Academic Medicine*, 75, p. S1517.

AMOS, D.E. & MASSAGLI, T.L. (1996) Medical school achievements as predictors of performance in a physical medicine and rehabilitation residency, *Academic Medicine*, 71, pp. 678–680.

ARNOLD, L. & WILLOUGHBY, T.L. (1993) The empirical association between student and resident physician performances, *Academic Medicine*, 68, pp. S35–S40.

BELL, J.G., KANELLITSAS, I. & SHAFFER, L. (2002) Selection of obstetrics and gynecology residents on the basis of medical school performance, *Am J Obstet Gynecol*, 186, pp. 1091–1094.

BEST EVIDENCE MEDICAL EDUCATION, CENTRE FOR MEDICAL EDUCATION, DUNDEE, UK (2002) [Online]. Available at: http://www.bemecollaboration.org

BLACKLOW, R.S., GOEPP, C.E. & HOJAT, M. (1993) Further psychometric evaluations of a class–ranking model as a predictor of graduates' clinical competence in the first year of residency, *Academic Medicine*, 68, pp. 295–297.

BOROWITZ, S.M., SAULSBURY, F.T. & WILSON, W.G. (2000) Information collected during the residency match process does not predict clinical performance, *Archives of Pediatric and Adolescent Medicine*, 154, pp. 256–260.

BOYSE, T.D., PATTERSON, S.K., COHAN, R.H., KOROBKIN, M., FITZGERALD, J.T., OH, M.S. *et al.* (2002) Does medical school performance predict radiology resident performance?, *Academic Radiology*, 9, pp. 437–445.

BRAILOVSKY, C., CHARLIN, B., BEAUSOLEIL, S., COTE, S. & VAN DER VLEUTEN, C. (2001) Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test, *Medical Education*, 35, pp. 430–436.

CALLAHAN, C.A., ERDMANN, J.B., HOJAT, M., VELOSKI, J.J., RATTNER, S., NASCA, T.J. & GONNELLA, J.S. (2000) Validity of faculty ratings of students' clinical competence in core clerkships in relation to scores on licensing examinations and supervisors' ratings in residency', *Academic Medicine*, 75, pp. S71–73.

CASE, S.M. & SWANSON, D.B. (1993) Validity of NBME Part I and Part II scores for selection of residents in orthopedic surgery, dermatology and preventive medicine, *Academic Medicine*, 68, pp. S51–S56.

CLUTE, K.F. (1963) *The General Practitioner: A Study of Medical Education and Practice in Ontario and Nova Scotia* (University of Toronto Press, Toronto).

COMPREHENSIVE META-ANALYSIS VERSION 1.0.23 BIOSTAT. 1998. [Online]. Available at: http://www.meta-analysis.com

COX, K. (2000) Examining and recording clinical performance: a critique and some recommendations, *Education for Health*, 13, pp. 45–52.

DONABEDIAN, A. (1998) The quality of care, *Journal of the American medical Association*, 260, pp. 1743–1748.

Dreyfus, S.E. & Dreyfus, H.L. (2001) A five stage model of the mental activities involved in directed skill acquisition. Unpublished manuscript supported by the Air Force Office of Scientific Research under contract F49620-79-C-0063 with the University of California, Berkeley.

Epstein, R.M. & Hundert, E.M. (2002) 'Defining and Assessing Professional Competence', *Journal of the American Medical Association*, 287, pp. 226–235.

Erlandson, E.E., Calhoun, J.G., Barrack, F.M., Hull, A.L., Youmans, L.C., Davis, W.K. & Bartlett, R.H. (1982) Resident selection: applicant selection criteria compared with performance, *Surgery*, 92, pp. 270–275.

Executive Council, World Federation for Medical Education (1998) International standards in medical education: assessment and accreditation of medical schools' educational programs. A WFME position paper, *Medical Education*, 32, pp. 549–558.

Ferguson, E., James, D. & Madeley, L. (2002) Factors associated with success in medical school: systematic review of the literature, *British Medical Journal*, 324, pp. 952–957.

Fincher, R.-M.E., Lewis, L.A. & Kuske, T.T. (1993) Relationships of interns' performances to their self-assessments of their preparedness for internship and to their academic performances in medical school, *Academic Medicine*, 68, pp. S47–S50.

Fine, P.L. & Hayward, R.A. (1995) Do the criteria of resident selection committees predict residents' performances?, *Academic Medicine*, 70, pp. S34–S38.

Fish, D.E., Radfar-Baublitz, L.S., 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, pp. 686–691.

General Medical Council (1993) *Tomorrow's Doctors*: *Recommendations on Undergraduate Medical Education*. [Online]. Available at http://www.gmc-uk.org/med_ed/tomdoc.htm

Gonnella, J.S. & Hojat, M. (1983) Relationship between performance in medical school and postgraduate competence, *Medical Education*, 58, pp. 679–685.

Gonnella, J.S., Hojat, M., Erdmann, J.B. & Veloski, J.J. (1993) A case of mistaken identity: Signal and noise in connecting performance assessments before and after graduation from medical school, *Academic Medicine*, 68, pp. S9–16.

Gonnella, J.S., Erdmann, J.B. & Hojat, M. (2004) An empirical study of the predictive validity of number grades in medical school using 3 decades of longitudinal data: implications for a grading system, *Medical Education*, 38, pp. 425–434.

Gunzburger, L.K., Fraizier, R.G., Yang, L.M., Rainey, M.L. & Wronski, T. (1987) Premedical and medical school performance in predicting first-year residency performance, *Journal of Medical Education*, 62, pp. 379–384.

Haig, A. & Dozier, M. (2003) BEME Guide No 3: Systematic searching for evidence in medical education—Part 1: sources of information, *Medical Teacher*, 25, pp. 352–363.

Harden, R.M., Grant, J., Buckley, G. & Hart, I.R. (1999) BEME Guide No 1: Best Evidence Medical Education, *Medical Teacher*, 21, pp. 3–15.

Heneman, R.L. (1983) The effects of time delay in rating and amount of information observed on performance rating accuracy, *Academy of Management Journal*, 26, pp. 677–686.

Hermen, M.W., Velowski, J.J. & Hojat, M. (1983) Validity and importance of low ratings given to medical graduates in non-cognitive areas, *Journal of Medical Education*, 58, pp. 837–843.

Hojat, M., Gonnella, J.S., Erdmann, J.B. & Veloski, J.J. (1997) The fate of medical students with different levels of knowledge: are the basic sciences relevant to physician competence, *Advances in Health Sciences Education*, 1, pp. 179–196.

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

Hojat, M., Veloski, J.J. & Borenstein, B.D. (1986) Components of clinical competence ratings: an empirical approach', *Education Psychology Measurements*, 46, pp. 761–769.

Holmboe, E.S. & Hawkins, R.E. (1998) Methods of evaluating the clinical competence of residents in internal medicine: a review, *Annals of Internal Medicine*, 129, pp. 42–48.

Kahn, M.J., Merrill, W.W., Anderson, D.S. & Szerlip, H.M. (2001) Residency program director evaluations do not correlate with performance on a required 4th-year objective structured clinical examination, *Teaching & Learning in Medicine*, 13, pp. 9–12.

Kane, M.T. (1992) The assessment of professional competence, *Evaluation & the Health Professions*, 15, pp. 163–182.

Kirkpatrick, D.I. (1967) Evaluation of Training, in: R. Craig & I. Mittel (Eds) *Training and Development Handbook*, (New York, McGraw Hill).

Korman, M. & Stubblefield, R.L. (1971) Medical schools evaluation and internship performance, *Journal of Medical Education*, 46, pp. 670–673.

Kron, I.L., Kaiser, D.L., Nolan, S.P., Rudolf, L.E., Muller, W.H. & Jones, R.S. (1985) Can success in the surgical residence be predicted from pre–residency evaluation?, *Annals of Surgery*, 202, pp. 694–695.

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

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

Martin, J.A., Reznick, R.K., Rothman, A, Tamblyn, R.M. & Regehr, G. (1996) BEME Systematic Review: Predictive values of measurements obtained in medical schools and future performance in medical practice, *Medical Teacher*, 71, pp. 170–175.

McManus, I.C., Smithers, E., Partridge, P., Keeling, A. & Fleming, P.R. (2003) A levels and intelligence as predictors of medical careers in UK doctors: 20 year prospective study, *British Medical Journal*, 327, pp. 139–142.

METRO Project 2004. [Online]. Available at: http://srv1.mvm.ed.ac.uk/metro/index/asp (accessed 7 June 2004).

Miller, G.E. (1990) The assessment of clinical skills/competence/performance, *Academic Medicine*, 65, pp. S63–S67.

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

Papadakis, M.A., Hodgson, C.S., Teherani, A. & Kohatsu, N.D. (2004) Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board, *Academic Medicine*, 79, pp. 244–249.

Patterson, F., Ferguson, E., Lane, P., Farrell, K., Martlew, J. & Wells, A. (2000) A competency model for general practice: implications for selection, training and development, *British Journal of General Practice*, 50, pp. 188–193.

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

Peterson, O.L., Andrews, L.P., Spain, R.S. & Greenberg, B.G. (1956) An analytical study of North Carolina general practice, Part 2, *Journal of Medical Education*, 31, pp. 1–165.

Price, P.B. (1969) Search for excellence, *American Journal of Surgery*, 118, pp. 815–821.

Price, P.B., Taylor, C.W., Nelson, D.E., Lewis, E.G., Loughmiller, G.C., Mathiesen, R. *et al.* (1973) *Measurement and Predictors of Physician Performance: Two Decades of Intermittently Sustained Research* (Salt Lake City, UT Aaron Press).

Price, P.B., Taylor, C.W., Richards, J.M. & Jacobsen, T.C. (1964) Measurement of Physician Performance, *Journal of Medical Education*, 39, pp. 203–211.

Prideaux, D., Alexander, H., Bower, A., Dacre, J., Haist, S., Holly, B. *et al.* (2000) Clinical teaching: maintaining an educational role for

doctors in the new health care environment, *Medical Education*, 34, pp. 820–826.

PROBERT, C.S., CAHILL, D.J., MCCANN, G.L. & BEN-SHLOMO, Y. (2003) Traditional finals and OSCEs in predicting consultant and self-reported clinical skills of PRHOs: a pilot study, *Medical Education*, 37, pp. 597–602.

RABINOWITZ, H.K. & HOJAT, M.A. (1989) Comparison of the modified essay question and multiple choice question formats: their relationship to clinical performance, *Family Medicine*, 21, pp. 364–367.

RAM, P. (1998) Comprehensive assessment of general practitioners: a study on validity, reliability and feasibility. Thesis, Maastricht.

RENGER, R. & MEADOWS, L. (1994) Testing for predictive validity in health care education research: a critical review, *Academic Medicine*, 69, pp. 685–687.

RETHANS, J.-J. (1991) *Does Competence Predict Performance? Standardized Patients as a Means to Investigate the Relationship between Competence and Performance of General Practitioners*. Thesis. (Amsterdam: Thesis) Publishers.

RICHARDS, J.M., TAYLOR, C.W. & PRICE, P.B. (1962) The prediction of medical intern performance, *Journal of Applied Psychology*, 46, pp. 142–146.

ROLFE, I.E., ANDREN, J.M., PEARSON, S., HENSLEY, M.K. & GORDON, J.J. (1995) Clinical competence of interns in New South Wales, Australia, *Medical Education*, 29, pp. 225–230.

RONAI, A.K., GOLMON, M.E., SHANKS, C.A., SCHAFER, M.F. & BRUNNER, E.A. (1984) Relationship between past academic performance and results of specialty in-training examinations, *Journal of Medical Education*, 59, pp. 341–344.

ROYAL COLLEGE of SURGEONS OF EDINBURGH (1997) *Guidelines for Basic Surgical Training* (Edinburgh, Royal College of Surgeons of Edinburgh).

RUTALA, P.J., FULGINITI, J.V., MCGEAGH, A.M., LEKO, E.O., KOFF, N.A. & WITZKE, D.B. (1992) Validity studies using standardized-patient examinations: standardized patient potpourri, *Academic Medicine*, 67, pp. S60–S62.

SCHMIDT, H., NORMAN, G. & BOSHUIZEN, H.A. (1990) Cognitive perspective on medical expertise: theory and implications, *Academic Medicine*, 65, pp. 611–621.

SMITH, S.R. (1993) Correlations between graduates' performances as first-year residents and their performances as medical students, *Academic Medicine*, 68, pp. 633–634.

SOSENKO, J., STEKEL, K.W., SOTO, R. & GELBARD, M. (1993) NBME Examination Part I as a predictor of clinical and ABIM certifying examination performances, *Journal of General Internal Medicine*, 8, pp. 86–88.

SOUTHGATE, L., HAYS, R.B., NORCINI, J., MULHOLLAND, H., AYERS, B., WOLLISCROFT, J. *et al.* (2001) Setting performance standards for medical practice: a theoretical framework, *Medical Education*, 35, pp. 474–481.

STEVENS, D.P. (2000) Three questions for the LCME, *Academic Medicine*, 75, pp. 960–961.

TAMBLYN, R., ABRAHAMOQICZ, M., DAUPHINEE, W.D., HANLEY, J.A., NORICINI, J., GIRARD, H. *et al.* (2002) Association between licensure examination scores and practice in primary care, *Journal of the American Medical Association*, 288, pp. 3019–3026.

TAMBLYN, R.M. (1994) Is the public being protected? Prevention of suboptimal medical practice through training programs and credentialing examinations, *Evaluating Health Professions*, 17, pp. 198–221.

TAMBLYN, R.M., BENAROYA, S., SNELL, L., MCLEOD, P., SCHNARCH, B. & ABRAHAMOWICZ, M. (1994) The feasibility and value of using patient satisfaction ratings to evaluate internal medicine residents, *Journal of General Internal Medicine*, 9, pp. 146–152.

TAYLOR, C.W. & ALBO, D. (1993) Measuring and predicting the performances of practicing physicians: an overview of two decades of research at the University of Utah, *Academic Medicine*, 68, pp. S65–S67.

VAN DER VLEUTEN, C. (2000) Validity of final examinations in undergraduate medical training, *Bahrain Medical Journal*, 321, pp. 1217–1219.

VISWESVARAN, C., ONES, D. & SCHMIDT, F. (1996) Comparative analysis of the reliability of job performance ratings, *Journal of Applied Psychology*, 81, pp. 557–574.

VU, N.V., BARROWS, H.S., MARCY, M.L., VERHULST, S.J., COLLIVER, J.A. & TRAVIS, T. (1992) Six years of comprehensive, clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine, *Academic Medicine*, 67, pp. 42–50.

VU, N.V., DISTLEHORST, L.H., VERHULST, S.J. & COLLIVER, J.A. (1993) Clinical performance-based test sensitivity and specificity in predicting first-year residency performance', *Academic Medicine*, 68, pp. S41–S45.

WASS, V., VAN DER VLEUTEN, C., SHATZER, J. & JONES, R. (2001) Assessment of clinical competence, *Lancet*, 357, pp. 945–949.

WEST, P.A. (2001) Calibre of people recruited to medicine may be too high for the job [letter], *British Medical Journal*, 322, p. 1361.

WILKINSON, T.J. & FRAMPTON, C.M. (2004) Comprehensive undergraduate medical assessments improve prediction of clinical performance, *Medical Education*, 38, pp. 1111–1116.

WILLIAMS, R.G., KLAMEN, D.A. & MCGAGHIE, W.C. (2003) Cognitive, social and environmental sources of bias in clinical performance ratings, *Teaching and Learning in Medicine*, 15, pp. 270–292.

WINGARD, J.R. & WILLIAMSON, J.W. (1973) Grades as predictors of physicians career performance: an evaluative literature review, *Medical Education*, 48, pp. 311–322.

WOLF, F.M., SHEA, J.A. & ALBANESE, M.A. (2001) Toward setting a research agenda for systematic reviews of evidence of the effects of medical education, *Teaching and Learning in Medicine*, 13, pp. 54–60.

WORLD FEDERATION FOR MEDICAL EDUCATION (2003) Postgraduate Medical Education. *WFME Global Standards for Quality Improvement* (Copenhagen, WFME). [Online]. Available at: http://www.wfme.org

YINDRA, K.J., ROSENFIELD, P.S. & DONNELLY, M.B. (1998) Medical school achievements as predictors of residency performance, *Journal Medical Education*, 63, pp. 356–363.

ZU, G., VELOSKI, J.J. & HOJAT, M. (1998) Board certification: associations with physicians' demographics and performances during medical school and residency, *Academic Medicine*, 73, pp. 1283–1289.