2008; 30: 124–145 MEDICAL TEACHER

BEME GUIDE

The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10

IAIN COLTHART[1], GELLISSE BAGNALL[1], ALISON EVANS[2], HELEN ALLBUTT[1], ALEX HAIG[1], JAN ILLING[3] & BRIAN MCKINSTRY[4]

[1]NHS Education for Scotland, [2]University of Leeds, [3]University of Newcastle, [4]University of Edinburgh, UK

Abstract

Review date: Literature search January 1990 to February 2005 (with update February 2006). Analysis completed January 2007.

Background: Health professionals are increasingly expected to identify their own learning needs through a process of ongoing self-assessment. Self-assessment is integral to many appraisal systems and has been espoused as an important aspect of personal professional behaviour by several regulatory bodies and those developing learning outcomes for clinical students.

In this review we considered the evidence base on self-assessment since Gordon's comprehensive review in 1991.

The overall aim of the present review was to determine whether specific methods of self-assessment lead to change in learning behaviour or clinical practice.

Specific objectives sought evidence for effectiveness of self-assessment interventions to:

a. improve perception of learning needs;
b. promote change in learning activity;
c. improve clinical practice;
d. improve patient outcomes.

Methods: The methods for this review were developed and refined in a series of workshops with input from an expert BEME systematic reviewer, and followed BEME guidance. Databases searched included Medline, CINAHL, BNI, Embase, EBM Collection, Psychlit, HMIC, ERIC, BEI, TIMElit and RDRB. Papers addressing self-assessment in all professions in clinical practice were included, covering under- and post-graduate education, with outcomes classified using an extended version of Kirkpatrick's hierarchy. In addition we included outcome measures of accuracy of self-assessment and factors influencing it. 5,798 papers were retrieved, 194 abstracts were identified as potentially relevant and 103 papers coded independently by pairs using an electronic coding sheet adapted from the standard BEME form. This total included 12 papers identified by hand-searches, grey literature, cited references and updating. The identification of a further 12 papers during the writing-up process resulted in a total of 77 papers for final analysis.

Results: Although a large number of papers resulted from our original search only a small proportion of these were of sufficient academic rigour to be included in our review. The majority of these focused on judging the accuracy of self-assessment against some external standard, which raises questions about assumed reliability and validity of this 'gold standard'. No papers were found which satisfied Kirkpatrick's hierarchy above level 2, or which looked at the association between self-assessment and resulting changes in either clinical practice or patient outcomes.

Thus our review was largely unable to answer the specific research questions and provide a solid evidence base for effective self-assessment.

Despite this, there was some evidence that the accuracy of self-assessment can be enhanced by feedback, particularly video and verbal, and by providing explicit assessment criteria and benchmarking guidance. There was also some evidence that the least competent are also the least able to self-assess accurately. Our review recommends that these areas merit future systematic research to further our understanding of self-assessment.

Conclusion: As in other BEME reviews, the methodological issues emerging from this review indicate a need for more rigorous study designs. In addition, it highlights the need to consider the potential for combining qualitative and quantitative data to further our understanding of how self-assessment can improve learning and professional clinical practice.

Correspondence: Dr Brian McKinstry, General Practice Section, Division of Community Health Sciences, The University of Edinburgh, 20 West Richmond Street, Edinburgh, EH8 9DX, UK. Tel: (0131) 650 2683; fax: (0131) 650 9119; email: Brian.McKinstry@ed.ac.uk

ISSN 0142–159X print/ISSN 1466–187X online/08/020124–22 © 2008 Informa UK Ltd.
DOI: 10.1080/01421590701881699

Practice points

- There is no solid evidence base within the health professions' literature which establishes the effectiveness of self-assessment in: identifying learner needs; influencing learning activity; changing clinical practice.
- The accuracy of self-assessment in clinical training may be improved by increasing the learner's awareness of the standard to be achieved.
- There is some indication that practical skills in clinical training may be better self-assessed than knowledge-based activities.
- Self-assessment needs to be used as one tool amongst other sources of feedback to provide a more complete appraisal of competence in health care practice.
- Future research should address the role that self-assessment plays in the everyday practice of health care decision-making.

Introduction

Health professionals are expected to identify their own learning needs through a process of ongoing self-assessment. Self-assessment is integral to most appraisal systems (British Medical Association 2003) and has been espoused as an important aspect of personal professional behaviour by several regulatory bodies (American Medical Association 1992; United Kingdom Central Council for Nursing, Midwifery and Health Visiting 1999; French Medicine Association 2002) and those developing learning outcomes for clinical students (General Medical Council 2002).

However there is no universal agreement on what constitutes self-assessment and its value in professional development is controversial (Eva & Regehr 2005). At one end of the spectrum, self-assessment is perceived as a quantifiable ability to predict individual performance in terms of an objective assessment measure, such as a multiple choice questionnaire. At the other end of the spectrum, it has been viewed as part of identifying everyday learning needs in the context of good professional practice. Although previous reviews of the literature on self-assessment (Gordon 1991, 1992; Ward et al. 2002) have suggested that the ability to self-assess is often lacking, the paucity of high quality research in this area raises questions about such conclusions.

In this review we consider the evidence base on self-assessment since Gordon's comprehensive review in 1991, in particular to determine whether specific methods of self-assessment in a clinical education context lead to change in learning activity or clinical practice.

Definitions of self-assessment

Self-assessment has been defined in a variety of ways, and the literature was consulted to inform the operational definition for this review.

Gordon (1991) suggests that the process of professionalisation should "provide the trainee with norms and expectations of professional behaviour, including recognition of one's own abilities and limitations". He defines *valid* self-assessment as "judging one's performance against appropriate criteria", and *accurate* self-assessment as "gaining reasonable concurrence between self-claimed and other, validated measures of performance".

Boud (1995) also addresses the importance of appropriate criteria for judging one's own performance, and emphasises the need for assessment standards and criteria to be made explicit. He defines self-assessment as "the involvement of students in identifying standards and/or criteria to apply to their work and making judgements about the extent to which they have met these criteria and standards".

Ward et al. (2002) implies that self-assessment is the "ability to accurately assess one's strengths and weaknesses", and follows on from Gordon (1992) in suggesting that this ability is "critical to the enterprise of lifelong learning".

On the basis of the above literature available to us in advance of the systematic review, and after much debate, we agreed an operational definition of self-assessment for this review:

> A personal evaluation of one's professional attributes and abilities against perceived norms.

Inclusion and exclusion criteria

It was agreed that in order to address the aim of our review, the focus of interest would be on interventions to aid this personal evaluation, and we thus excluded papers where there was no description of an explicit self-assessment tool or method. We therefore excluded unstructured self-reflection as a self-assessment intervention.

In debating our definition of self-assessment, and in refining our inclusion criteria, we were conscious that the concept of self-assessment generally implies an element of individual introspection, and thus inevitably overlaps with the psychological literature on self-referent thinking, which is recognized as a "key variable in clinical, educational, social, developmental, health and personality psychology" (Schwarzer 2005). Woolliscroft et al. (1993) argued that self-assessment is "central to the function of the clinician" (p. 290) and they define self-assessment in terms of a "self representation of actual performance". This issue will be addressed more fully in the discussion section. However, it is important to note here that we carefully considered whether to include papers relating to self-efficacy. This concept generally refers to a person's judgements about his/her abilities to deal with their experiences. Bandura (1982) argued that self-efficacy influences what people choose to do, whether they approach tasks with anxiety or confidence, how much effort they devote to tasks, and how long they persist in the face of disappointment. More recently, Bandura (1994) defined "perceived self-efficacy as people's beliefs about their capabilities to produce designated levels of performance that exercise influence over events that affect their lives".

In line with our definition of self-assessment, we thus decided to include papers on self-efficacy only if these included the use of a self-assessment tool or explicit method.

We also had to make a decision about clinical audit, which can be an effective indicator of quality of performance and in the widest sense could be regarded as a self-assessment tool. The effectiveness of audit used in this way, however, is established (Jamtvedt et al. 2006) and so we agreed to exclude papers that exclusively describe audit systems per se. We decided that we would include studies that use audit as a means of establishing the effectiveness of self-assessment.

Papers published since we began our review, e.g. Eva and Regehr (2005), have expanded our thinking around definitions of self-assessment. These issues will be explored in the discussion section of the paper, including the distinction between self-efficacy and self-concept.

Previous research

Gordon (1991, 1992)

In his 1991 review of the validity and accuracy of self-assessments in health professions training, Gordon identified some useful consistencies in findings regarding self-assessment. His main findings are noted below as a background to the current review. However the need for updating this review is indicated by his comments regarding diverse theoretical backgrounds to studies, little continuity and absence of rigorous research methods.

Gordon classified studies into four categories:

(1) Experiments in which self-claimed factual knowledge was tested against verifiable facts.

These studies seemed to show a tendency towards overconfidence, particularly amongst those students who knew less.

(2) Studies in which health professions trainees viewed samples of their own clinical behaviour on videotape and assessed their performances using behavioural rating instruments.

In the four studies in this category, student and faculty ratings of clinical skills were compared. Higher correlation coefficients were associated with more specific clinical tasks.

(3) Global self-assessments of performance based on extended periods of supervized functioning in clinical training environments.

Gordon concluded that the findings supported "the hypothesis that self-assessments are strongly influenced by global self-attributions and are perhaps as closely linked to self-concept as they are to previous performance".

(4) Studies of innovative training programs in which valid and accurate self-assessment was an explicit goal and in which specific strategies for improving self-assessment skills were used.

Five studies included here showed that "students' self-assessment skills were improved by clarifying the criteria for success, by reconciling self-assessments with supervisors' judgements or other external performance measures, and by linking accurate self-assessment to success or increased student control in the course".

Gordon concluded from this 1991 review that "self-assessment skills remain underdeveloped during training". In his 1992 review of self-assessment programmes, he describes two common characteristics of effective programs to improve the validity and accuracy of self-assessment:

- an expectation that learners would systematically gather and interpret data on their performances;
- formal requirements to reconcile learners' self-assessments with credible external evaluation sources.

He found a diverse theoretical background to the research studies, and no indication that later studies built on the advances of earlier ones. The eleven studies he reviewed were "not scientific experiments with rigorous designs, but modest attempts at curricular innovation". There did seem to be consistency in their findings, however. Students may at first be uncomfortable with the concept of self-assessment, and not trust their tutors or school sufficiently to assess themselves honestly. Programmes that successfully made the transition to enthusiastic student participation in self-assessment had strong student representation in their planning, explicit rules on confidentiality, and "patience in winning the confidence of the residents".

These conclusions could have important implications for clinical education, but key recommendations around self-assessment in clinical education would need to be based on more robust evidence.

Kruger and Dunning (1999)

Important insights as to why self-assessment might be inaccurate and how self-assessment skills might be improved were gained from Kruger and Dunning's (1999) paper. They set out to test in a non-clinical context the hypotheses that "incompetent individuals have more difficulty recognizing their true level of ability than do more competent individuals, and that a lack of meta-cognitive skills may underlie this deficiency". In three different areas of testing (humour, logical reasoning, and grammar) they found that those who scored in the bottom quartile grossly overestimated their own abilities, both with respect to their peers and in estimating their actual scores. Those who scored in the top quartile tended to underestimate their performance, but were still more accurate in their self-assessments than those in the bottom quartile. However these results could also be explained by a regression towards the mean.

Those in the top and bottom quartiles on a grammar test were asked to re-rate their own performances after benchmarking. This was carried out by asking them to rate the performances of five participants whose results had the same range as the overall population of participants. Those in the bottom quartile were less able to accurately rate the peer performances than were those in the top quartile. In addition, those in the bottom quartile slightly increased their own already inflated self-ratings after this exercise, making them even more inaccurate, whereas those in the top quartile also raised their self-assessment estimates, making them more accurate.

The conclusion here was that the "incompetent individuals fail to gain insight into their own incompetence by observing the behaviour of other people". It also seemed that a 'false-consensus effect' had been operating with the high scorers assuming that their peers would also be high scorers. Seeing a range of performances helped the high scorers to re-calibrate themselves in relation to their peers.

A fourth study in this series tested the prediction that the meta-cognitive skills of the poor performers could be improved by giving them training to make them more competent in logical reasoning, and thus providing them with the meta-cognitive skills necessary to be able to realize that they have performed poorly. Following training in logical reasoning, the low scorers on this test improved both their logical reasoning skills and their accuracy in self-rating to be nearly as accurate as the high scorers. Further analyses of the results showed that it was the improved metacognitive skills that enabled the less competent to become more accurate in their self-assessment.

Previous research methods

Previous reviews (Gordon 1991; Ward et al. 2002) suggest that much of the evidence for poor accuracy of self-assessment was based on quantitative studies, some of which used group analyses to compare ratings of students and teachers, often with un-validated rating scales. Individual accuracy in identifying strengths and weaknesses would not be identified in such studies. These issues have been discussed at length by Ward et al. (2002) and will be explored in more detail later in the report.

For the reasons given above, it is unlikely that such studies will give us a complete picture of the accuracy and usefulness of self-assessment in the health professions. In this review, therefore, we have not limited ourselves to particular research methods, but have selected on the basis of study quality and whether the conclusions are important and likely to be applicable in contexts other than that of the original research.

As noted in the introduction, the importance of updating our understanding of self-assessment in clinical education is emphasised by the increasingly widespread assumption that learners will accurately identify their own learning needs through self-assessment.

Given that self-assessment is generally accepted as a pre-requisite for continuing professional development (CPD) in the health professions, our review question centred on the evidence around self-assessment interventions. In line with other Best Evidence Medical Education (BEME) reviews (Dornan et al. 2006; Hammick et al. 2007) we wanted to know if there was evidence of self-assessment interventions improving outcomes at each level of Kirkpatrick's hierarchy (Kirkpatrick 1967).

Objectives of the review

The following objectives were identified for the conduct of the review.

Box 1. Abbreviations.

AHP – Allied Health Professional
AMEE – Association for Medical Education in Europe
ASME – Association for the Study of Medical Education
BEI – British Education Index database
BMA – British Medical Association
BNI – British Nursing Index database
CASP - Critical Appraisal Skills Programme
CINAHL – Cumulative Index to Nursing and Allied Health Literature database
CME – continuing medical education
CPD – continuing professional development
EBM Collection – Evidence Based Medicine Collection database
EKG - electrocardiogram
Embase – Excerpta Medica database
ERIC – Educational Resource Information Center database
FacSeS – Faculty Self-efficacy Scale
GPA - grade-point average
GMC – General Medical Council
GRS – Global Rating Scale
HMIC – Health Management Information Consortium database
IM - impression management
ITE - in-training examination
MCAT – Medical College Admission Test
MCQ - multiple choice questionnaire
Medline – US National Library of Medicine bibliographic database
NBME – National Board of Medical Examiners
NES - NHS Education for Scotland
OCRS – Operative Component Rating Scale
OG – Obstetrics and Gynaecology
OSATS - Objective Structured Assessment of Technical Skills
OSCE – Objective Structured Clinical Examination
Ottawa – International Ottawa Conference on Medical Education
PDS - Paulhus Deception Scale
PPI – Personal Progress Inventory
PsychLit – Psychology Literature database
RDRB – Research and Development Resource Base database
SAM – self-assessment manual
SDE - self-deception enhancement
SP - standardized patient
TIMElit – Topics in Medical Education (literature) database
UKCC - United Kingdom Central Council for Nursing, Midwifery and Health Visiting

- Identify the scope of the research on the effectiveness of self-assessment methods
- Review the evidence of the impact of self-assessment methods on

i. identification of learning needs
ii. learning activity
iii. clinical practice

- Identify the perceived value of self-assessment to learners
- Make recommendations for further research and practice

Review questions

- Are there effective self-assessment interventions which:
  - improve the accuracy of learner perception of their learning needs?
  - promote an appropriate change in learner learning activity?
  - improve clinical practice?
  - improve patient outcomes?

Box 2. Coding sheet questions.

Does this study meet ALL the following
INCLUSION CRITERIA?
1. Is it about self-assessment?
2. Is it set in a clinical training context?
3. Does it have either:
   i) an evaluation of the self-assessment method or tool? OR
   ii) offer important information about attitudes towards/perceptions of self assessment? OR
   iii) is it a comparison study (measuring accuracy of self-assessment against some other assessment)? OR
   iv) does it describe an impact of self assessment on teachers and/or learners?

TYPES OF STUDY INCLUDED
Comparison study: self versus external
Factors affecting self-assessment
Impact of self-assessment
Methods to improve self-assessment
Perceptions of self-assessment
Self-assessment tools – validity and reliability
Teaching assessment
EXCLUSION CRITERIA
Not original research (e.g. review)
No assessment of intervention and/or its impact
Not a clinical context
Not self-assessment (e.g. audit)
Self-assessment used to evaluate another programme or intervention (blind tool)
No structured self-assessment method described

Subsidiary research questions:

- What are the factors affecting the accuracy of self-assessment in relation to other assessments such as peer and external?
- What are learners' and teachers' perceptions of and attitudes to self-assessment?

Review methodology

The methods for this review were developed and refined in a series of workshops with input from an expert BEME systematic reviewer, and followed BEME guidance. The research protocol was submitted to BEME for peer review.

Inclusion and exclusion criteria

The inclusion and exclusion criteria were drawn up in line with our definition of self-assessment to ensure that the papers selected would be relevant and focused on the research questions. The criteria are listed in Box 2. Although review papers were not used in answering our research questions, we have referred to relevant reviews in our discussion.

1. Types of studies – research designs

All research designs were considered (see Box 3). These categories were derived from the initial review of abstracts and reflect the content of the abstracts rather than formal theoretical frameworks within educational research. Many studies were not explicit about their underlying theoretical framework, and we wanted to ensure we could incorporate all relevant approaches.

We included studies that compared the accuracy of self-assessment in a variety of clinical settings with peer or tutor assessment in order to determine if particular groups of learners are more accurate than others in self-assessment. We also considered studies that explored the attitudes of learners and teachers to self-assessment. To help understand the range of methods employed within these research designs information was recorded on data collection methods (e.g. interviews, questionnaires, and observations) and analysis (qualitative, quantitative or both). We also recorded the type of clinical setting in which the intervention took place and the professional context involved. Finally we recorded synonyms and definitions of self-assessment used by different authors.

Box 3. Research designs.

Type of study	
Pilot	Prospective
Qualitative	Retrospective
Quantitative	Randomized trial
Single group study	Comparative
Cohort study	Action research
Case control	Case study
Cross sectional	Historical
Before and after study	Meta-analysis
Time series	Narrative
Non-randomized trial	

2. Types of self-assessment intervention

We considered all forms of structured self-assessment which included an explicit intervention method or tool. In addition we included studies of interventions to improve the effectiveness of self-assessment.

3. Types of participants

We included all professions in clinical practice including chiropodists/podiatrists, complementary therapists, dentists, dieticians, doctors, hygienists, psychologists, psychotherapists, midwives, nurses, pharmacists, physiotherapists, occupational therapists, radiographers and speech therapists. We also included clinical undergraduate students from these specialties.

4. Types of outcome measures

Outcome measures were based on an extended version of Kirkpatrick's (1967) model of outcomes at four levels as shown in Box 4. We also included outcome measures of accuracy of self-assessment and the factors influencing self-assessment. Additional predetermined and unintended outcomes were also accepted.

Search Strategy

A comprehensive literature search was conducted across all sources relevant to professional education in a clinical context.

The database search covered all relevant health as well as educational databases, and included: Medline, CINAHL, BNI, Embase, EBM Collection, Psychlit, HMIC, ERIC, BEI, TIMElit and RDRB. The strategies were designed and tested for maximum sensitivity to ensure no potentially relevant papers

Box 4. Kirkpatrick's Hierarchy adapted to self-assessment.

Level 1 - Reaction

These cover learners' views on the self-assessment experiences, its perceived usefulness, possible general positive and negative effects on learning, self- esteem, relationship with tutors and peers.

Level 2 - Modification of attitudes/perceptions

These outcomes relate to specific perceived changes in individuals in respect to their perceptions of knowledge and skill in the tested area, specific impact on personal self-esteem and relationships with tutors and peers.

Level 3 – Change in learning behaviour

Recorded change in learning behaviour as a result of a self-assessment intervention.

Level 4a - Behavioural change

Actual change in clinical practice as a result of a self-assessment exercise.

Level 4b - Change in patient outcomes

Any improvements in the health and well-being of patients/clients as a direct result of self-assessment intervention. Where possible objectively measured or self-reported patient/client outcomes will be used, such as: health status measures, disease incidence, duration or cure rates, mortality, complication rates, readmission rates, adherence rates, patient or family satisfaction, continuity of care.

were missed. The search ran from January 1990 to February 2005 and did not limit by language, geography, or research methodology. An updating search was conducted in January 2006 to include evidence published during the course of this group's analysis. The full (Medline) search strategy is outlined in the full BEME report (see www.bemecollaboration.org).

The results of the database search were augmented by further methods. A cited reference search was conducted on the core papers of relevance examining which papers these cited, and in turn which future papers referred back to the core. This is a method BEME has found very effective for retrieving relevant papers that imperfect educational descriptors within clinical databases fail to adequately describe (Haig & Dozier 2003). Grey literature (evidence not formally or commercially published) searches were also conducted along BEME methodology.

Finally, hand searches were conducted across the most relevant journals: *Academic Medicine*, *Medical Teacher*, *Medical Education*, *Nurse Education in Practice* and *Education for Primary Care*, as it is recognized that electronic indexing of clinical education terms and clinical educational journals was unreliable at times throughout that period. Titles suggesting a focus on self-assessment that had not already been identified were obtained for examination of abstract and if indicated full text. References in full text articles were explored for additional citations.

The original list of retrieved articles was visually scanned to determine whether they potentially fulfilled the research questions. From this list the abstracts were obtained. All abstracts were viewed by at least two group members to decide if a full text version of the article should be obtained. The full text article was obtained if the abstract suggested that the focus of the study was self-assessment or that a validated form of self-assessment was described as part of the study, that the study took place in either an undergraduate or postgraduate clinical education setting and that it did not meet the exclusion criteria. Where there was disagreement on the decision to obtain a full article a third reviewer reviewed the abstract and a majority decision was made. The process of the review is summarized in Figure 1. From here it can be seen that 77 papers were agreed for final analysis; of these 39 were not considered as adequate to be informative, 32 were, and an additional 6 papers were included for their relevance although they did not satisfy all our inclusion criteria (e.g. a review rather than primary research).

Data abstraction

A coding form was devised from the BEME standard version, containing sections to determine the strength and relevance of the study to the research questions, as well as the rigour of the study design itself. The latter sections were adapted from the NHS Critical Appraisal Skills Programme (CASP) tools, which are widely used critical appraisal instruments created to objectively evaluate specific research methodologies (http://www.phru.nhs.uk/casp/casp.htm). In addition an instrument to assess the quality of comparative studies was devised by the group. The checklists appear in the coding sheet.

The coding sheets were designed to permit consistency across the different qualitative and quantitative approaches to data collection. All members of the review team independently coded a selection of papers into the data abstraction sheets to validate the coding sheets for utility and completeness.

All full papers were then read by two group members, using the final version of the coding sheet. As the group was split between different sites across the United Kingdom (Edinburgh, Glasgow, Newcastle, Leeds and Birmingham), a web-based coding form was developed to enable geographically separated pairs to code and agree data

(http://134.36.210.98/cgi-bin/survey/survey/24). Papers which on full reading did not meet the inclusion requirements were rejected and the reasons recorded (see Table 2 in the full BEME report: www.bemecollaboration.org).

Abstracted data included a detailed checklist for the different types of research method employed. Reviewers were asked to rate;

- the appropriateness of the design of the study to answer the research questions posed;

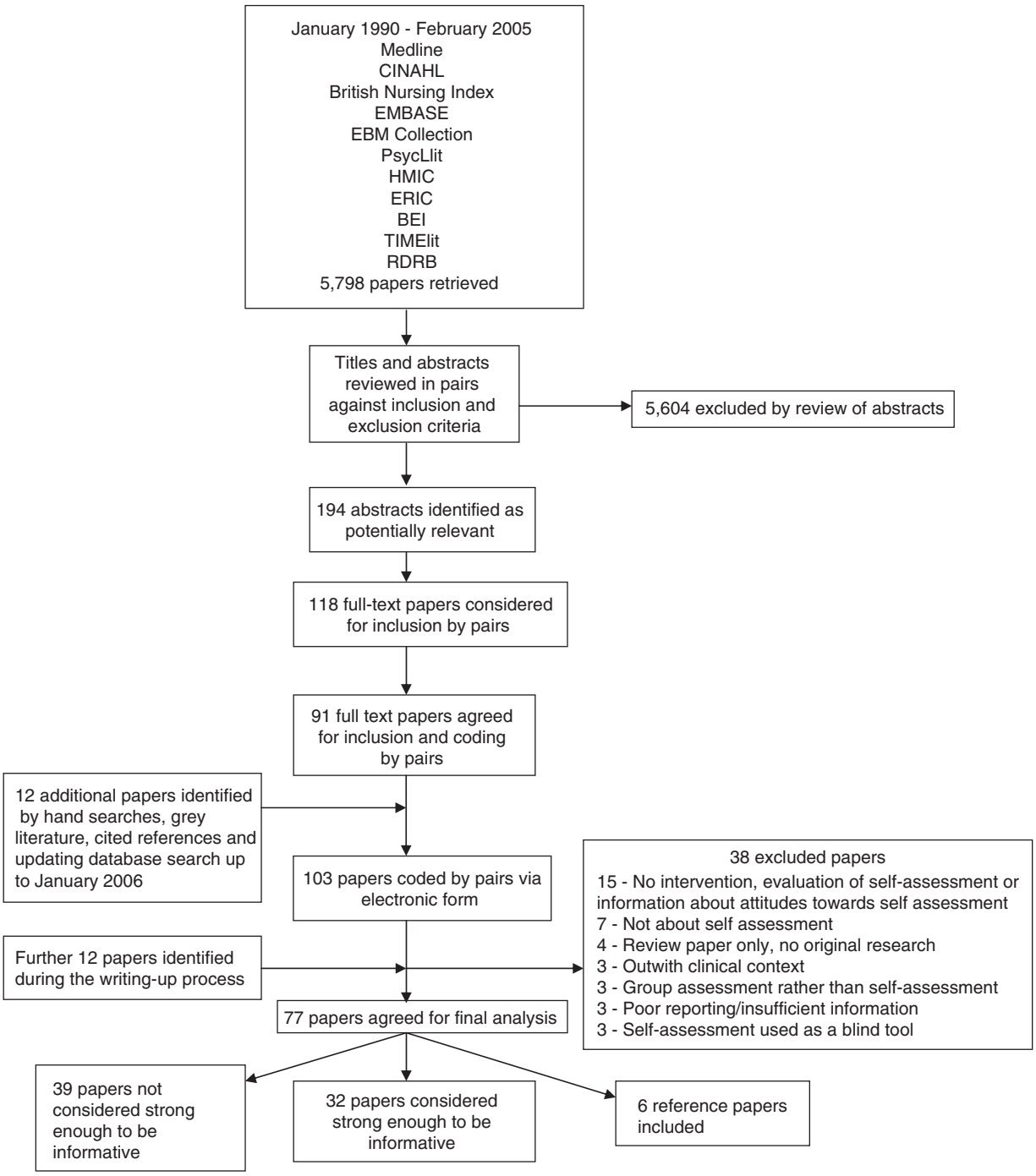

Figure 1. Flowchart of Search and Selection Strategy.

- how well the design was implemented;
- the appropriateness of the analysis;

and to comment on concerns. They were then asked to comment on what level of the Kirkpatrick Hierarchy (Kirkpatrick 1967) the outcomes related to. Additionally reviewers identified references cited in these papers that might be of interest to the review and where appropriate these were obtained.

Following data extraction of each paper the two group members independently scored them on a scale of 1 to 5 for the strength of the findings (Box 5a).

Papers where the conclusions were not supported by the evidence presented i.e. grades 1 and 2 were not considered further.

The perceived overall importance of the paper in terms of the rigour with which it was conducted, relevance, and generalisability was also graded independently by both reviewers (Box 5b). Again papers with grades 1 and 2 were discarded.

The reviewing pair then consulted and agreed final scores for the paper. As with the abstracts, any discrepancies were usually resolved through discussion between the pair. Inter-reviewer agreement was high, with adjudication being required on only three occasions.

Box 5a. Gradings of Strength of Findings of the Paper.

Grade 1 No clear conclusions can be drawn. Not significant.

Grade 2 Results ambiguous, but there appears to be a trend.

Grade 3 Conclusions can probably be based on the results.

Grade 4 Results are clear and very likely to be true.

Grade 5 Results are unequivocal.

Box 5b. Gradings of Overall Importance of the Paper.

Grade 1 Papers with numerous deficiencies in the rigour or appropriateness of the methodology or the statistical analysis.

Grade 2 Papers with some deficiencies in the rigour or appropriateness of the methodology or the statistical analysis.

Grade 3 Papers with doubts about the rigour or appropriateness of the methodology or the statistical analysis.

Grade 4 Papers with rigorous methodology and appropriate statistical analysis, but doubts about adequate sample size.

Grade 5 Papers with generalisable findings, rigorous methodology, adequate sample size, and appropriate statistical analysis.

Papers that scored 4 or above on either strength of findings or importance were considered to be higher quality papers and are reported fully in the review. All these papers were read again and summarized in an abbreviated format by three members of the team. 'Borderline' papers (rated 3 on strength of findings and on importance) were also reviewed independently to ensure that no higher quality paper had been excluded.

Analytical procedures–synthesising the findings

Although we were prepared if possible to undertake meta-analysis, we recognized that very few of the variables coded were likely to be ratio data, with some interval data. Most of the data were categorical and insufficiently homogeneous to allow meta-analysis of results. The review therefore was largely descriptive, with the results reported through a narrative framework that focused on key themes. These are summarized below and form the subheadings for reporting the results.

Key themes

- Peer Assessment and faculty ratings
- Individual characteristics
  - Gender
  - Cultural differences
  - Insight
- External factors
  - The purpose of the self-assessment task
  - Practical skills versus theoretical knowledge
- Factors influencing self-assessment
  - Video feedback and benchmarking
  - Video and verbal feedback
  - Instruction
  - Experience
- Perceptions and attitudes towards self-assessment

Each member of the review team undertook to synthesise data from papers that were considered to be of higher quality for one or more of the themes.

Results

Despite very inclusive strategies being employed (5,798 total hits were recorded) the conventional strategies were unable to retrieve all papers within the databases searched. The search specificity (the percentage of the returns that were actually relevant to the topic) was particularly poor at 3.3% and therefore time consuming for the group as thousands of false hits had to be discarded. This was due to ambiguities around searching for clinical education literature already researched by BEME (Haig and Dozier 2003), but also to the lack of clarity and consistency ascribed to the concept of self-assessment itself. Search sensitivity (the percentage of the total relevant papers retrieved) was also poor at 91%.

Although the search did not limit by geography or language, two thirds of the final papers were North American and over four fifths came from English-speaking countries. Homogeneity was also evident with regards to study design; while this group considered all research methods, less than 5% of included papers used only qualitative methods (Box 6).

Included papers are summarized in Table 1, and the excluded papers are listed in Table 2 with reasons for exclusion (see www.bemecollaboration.org for tables).

Methodological quality of studies

In many papers we reviewed, conventional good research practice was either not followed or the report of the study did not allow the reader to critically evaluate the study, as key pieces of information were not included. The review has identified a variety of such problems and these are outlined below.

- Assessment instruments used in some studies were either not validated or no reference was made to their reliability and validity
- There was a frequent assumption that expert opinion provided a gold standard, yet it was rare for validity or reliability of the expert opinions to be examined
- The use of group means in some comparison studies ignored individual variation in self-assessment ability
- In some studies control groups were needed but not used
- It was rare for power calculations to be provided. Few studies were set up to test specific hypotheses, and most were limited to correlational analyses
- Sampling and selection strategies were not stated in many studies, which meant that assessments could not be made of how representative the study participants were of their populations. Likewise many studies failed to present data

Box 6. Distribution of papers.

Geographical:	Education level:	Profession:
USA – 57%	Undergraduate – 76%	Medicine – 75%
UK – 14%	Postgraduate – 22%	Teaching Staff – 9%
Canada – 8%	CPD/CME – 2%	Dentistry – 7%
Australia – 4%		Nursing – 7%
Sweden – 3%		Allied Health Professionals – 1%
Others – 14%		Psychology – 1%

on non-participants, which casts doubt on the representativeness of the sample.
- Inadequate explanation of missing data
- Statistical methods unclear
- Study conducted at a single institution bringing into question the generalizability of the study
- No clear information presented on how qualitative data were analysed.

The aim of several papers was to correlate a self-assessed measure against an external measure. Typically the external measure was the judgement of an assessor (peer, faculty, tutor or clinical preceptor) or a criterion measure such as an examination or checklist. The validity and reliability of these external measures was rarely reported.

Specific research findings

This section reports the results from the 32 papers which scored 4 or above on either strength of findings or importance – i.e. 'high quality' papers.

Results are presented firstly in terms of

a. their ability to answer the original research questions for the review
b. themes which emerged from the papers. Each theme forms a subheading in section (b) below.

(a) Answers to research questions

Few papers treated self-assessment as an intervention in itself, and none of the high quality papers looked specifically for changes as a result of undertaking self-assessment alone.

Are there effective self-assessment interventions which:

(i) improve the accuracy of learner perception of their learning needs. The majority of the studies we found addressed the accuracy of self-assessment compared with an external assessment, but none of the high quality studies attempted to either measure change in perceptions of learning needs, or to find a valid assessment of learning needs against which to compare self-assessed needs. Interventions to improve the accuracy of self-assessment are discussed in a separate section below.

One paper that was difficult to classify did address the assessment of learning needs in children's hospice doctors (Amery & Lapwood 2004). This study was felt not to meet the inclusion criteria as there was no external comparator nor was there an evaluation of the self-assessment method. The findings, however, were interesting in that they highlighted the different learning needs identified when doctors completed questionnaires, and when they had an interview based on incidents reported in an educational diary. The authors suggest that a variety of methods are needed to fully identify learning needs, with 'self-perception analysis' being needed in addition to facilitation and diary keeping to help identify the areas that subjects don't know that they don't know.

(ii) promote appropriate change in learner learning activity – Kirkpatrick level 3. None of the high quality papers reported any self-assessment intervention that led to a change in learner's learning activity.

(iii) improve clinical practice/improve patient outcomes – Kirkpatrick level 4. Only two papers addressed this question:

Ericson et al. (1997) was recorded on the database as providing evaluation at level 4. The self-assessment exercise was carried out on 41 dental students and was accompanied by clinical guidelines, so it could be that the main educational effect was related to students following the guidelines rather than being the result of self-assessment. There was good agreement between tutors' and students' ratings (the same rating was given in 87% of instances, 10% of students under-rated themselves, and 3% over-rated). This study suggests that the use of guidelines might aid self-assessment, but there was no control group. It does not present any evidence that self-assessment on its own has any impact at any Kirkpatrick level.

The second paper recorded on the database as Kirkpatrick level 4 was Biernat et al. (2003). This study compared faculty assessments with residents' self-assessment skills of their performance in an interview with a simulated patient portraying dementia. Twelve residents undertook a videotaped interview then completed a checklist of behaviours carried out in the interview. The videotape was rated by a faculty member, then residents were able to review the tape with the programme director for feedback and additional instruction. The residents completed an evaluation form, all of them reporting that the self-assessment tool was useful (Kirkpatrick level 1). One comment indicated that the experience would change the way the resident treated patients with memory loss, and another reported being encouraged to improve knowledge (Kirkpatrick level 2). There was no test of whether the practice of the residents changed, or any measure of change in patient outcomes.

> In summary, we did not find any high quality papers to answer our main research questions, based on Kirkpatrick's hierarchy.

We did however find some useful evidence on our subsidiary research questions and on other themes relating

to self-assessment. Section (b) below summarizes the findings under sub-headings which reflect these themes. To facilitate interpretation, the text under each sub-heading includes a summary discussion. We hope that this will help the reader, rather than having all the comments in a separate discussion section, which would lead to repetition and difficulty in linking the findings with the relevant section of the discussion.

(b) Themes relating to self-assessment

Peer assessment and faculty ratings

A number of studies have specifically addressed the question of peer assessment in the context of self-assessment. Typically self-assessment was correlated against both peer ratings and expert opinion which may be represented by faculty or a tutor. The research suggests a consistent pattern of results in relation to how self-assessment rates against peer assessment. The following studies typify the general conclusion across a number of studies that individuals are more able to accurately assess their peers' ability than their own.

Rudy et al. (2001) compared self-assessment, peer and faculty evaluations of interviewing skills for 97 first year medical students. Although correlations were modest they found that individuals gave their peers a more balanced assessment in comparison to how they rated themselves. Correlations between self and peer ratings ($r=0.29$, df $=89$, $p=0.008$) and between faculty and peer ratings ($r=0.50$, df $=86$, $p=0.0001$) were statistically significant. The correlation between self and faculty composite scores showed marginal statistical significance ($r=0.19$, df $=80$, $p=0.08$). This leads them to conclude that students are capable of assessing their peers but have difficulty in accurately evaluating their own performance.

Sullivan et al. (1999) used a similar methodology by comparing self, peer and faculty ratings in the setting of a problem based tutorial group for 154 third year medical students.

They found that the medical students were not able to identify their own strengths and weaknesses as compared to their peers and faculty. Three areas were assessed in the context of the tutorial: independent learning, group participation and problem solving. Again correlations were moderate but they found the highest correlation between peer and faculty ratings: independent learning ($r=0.5$); group participation ($r=0.54$) and problem solving ($r=0.24$) (all significant at $p=0.01$). In comparison the lowest correlation was between self and faculty ratings: independent learning ($r=0.24$); group participation ($r=0.18$) and problem solving ($r=0.11$) (all significant at $p=0.05$).

Bryan et al. (2005) found that students received significantly more positive comments from their peers than from themselves. Students were also ranked higher by their peers than by themselves with a mean (±sd) of 4.3 (±0.5) and 3.6 (±0.8) respectively, $p<0.001$.

Rudy et al. (2001) also present a number of possible explanations why students are more proficient in evaluating their peers in comparison to their own skills, knowledge and performance. Firstly students may be socially uncomfortable in presenting a wholly favourable impression of themselves to others and prefer to be modest in their self-assessments. Alternatively students at a certain level of training may have unrealistic goals and expectations of their abilities due to inexperience. Another possible explanation is a tradition of judgemental and punitive evaluation in medical education which inhibits students from expressing themselves. The way individuals judge performances may also go some way to explaining this anomaly in that they assess their peers at face value but apply global perceptions of performance to their own abilities. Finally the method of self-assessment may influence the outcome. For example a study which uses video recording may contribute to inaccurate self-assessment by causing anxiety and self-consciousness.

The general consensus here (albeit limited to three studies) that individuals are more able to accurately assess their peer's performance in comparison to their own is valuable when considering methods of validating self-assessment. The triangulation of a self-assessment measure by a more accurate measure should increase the value and meaningfulness of the exercise for an individual.

Individual characteristics

A common aim of many studies was to identify factors and characteristics in individuals which would account for their differential ability to self-assess. There are two recurring themes which dominate the literature reviewed, namely gender and insight. There have been limited attempts to investigate the effects of cultural differences. Insight has become a field of study in itself as exemplified by the previously discussed work of Kruger & Dunning (1999). There is a separate section later in this section specifically addressing insight. With reference to Kruger & Dunning (1999) insight may be defined as the ability to assess how well one is performing, when one is likely to be accurate in judgment and when one is likely to be in error. Experience is also considered later under the heading 'Factors influencing self-assessment'. Gender and cultural differences in self-assessment are discussed below from papers included in our review.

Gender

Researchers consider gender an obvious starting point in looking for potential reasons for differences in outcomes when individuals self-assess. There are more papers reporting differences in gender than any other type of sub-analyses. Despite this, the evidence drawn from across a number of studies is either inconclusive or contradictory.

Edwards et al. (2003) intentionally set out to investigate the influence of demographic factors on the accuracy of self-assessment. Given its clear objective to assess the influence of gender differences, and the sample size of the study (1,152 students over a 10 year period) the results of this study deserve credence. It was found in the study population of third year medical students in an obstetrics and gynaecology clerkship that men were 1.7 times (odds ratio 1.72: 95% CI 1.53 to 1.93) more likely than women to overestimate their grades.

A similar conclusion was reached by Minter et al. (2005) who examined gender differences in surgical residents. The sample size was small (female n10, male n19) but nevertheless the authors found that both male and female residents underestimated their abilities compared with faculty. In comparison female residents underestimated their abilities to a greater extent (-1.15 ± 0.42 points) than their male counterparts (-0.75 ± 0.19 points) but the difference between the two groups was not significant.

Bryan et al. (2005) in a study of 213 medical students found that males rated themselves more highly than females (mean, ±sd) 3.7 ($= \pm 0.8$) and 3.5 ($= \pm 0.9$) respectively ($p = 0.04$). Males received significantly more positive comments than females on peer evaluations of 9.1 (±2.5) and 8.4 (±2.0) respectively ($p = 0.025$) and were rated higher than females on peer provided numerical rating (mean, ±sd) of 4.4 (±0.5) and 4.2 (±0.5) respectively ($p = 0.02$).

In contrast, Leopold et al. (2005) discovered contradictory evidence on gender differences in confidence levels depending on when the measure was taken. They examined the confidence and self-assessment of 93 practitioners in performing a simulated knee joint injection. Measures of confidence and self-assessment were taken before and after they were randomized to three types of instruction: printed manual; video; hands-on instruction. The self-assessment was compared with objective performance standards measured by a custom designed knee model with electronic sensors that detected correct needle placement. Prior to instruction male participants were significantly more confident (6.32 points on a 10 point Likert scale) than female participants (2.95 points, $p < 0.01$). In terms of performance there was no significant difference between the performances of men and women (6.62 and 5.86 points respectively, $p > 0.05$). After instruction female participants were significantly more confident than male participants (8.77 compared to 6.98 points, $p < 0.01$) and also had higher objective scores for performance (8.88 compared with 7.73 points, $p < 0.05$).

Zonia & Stommel (2000) compared 73 interns' self-assessments of their medical knowledge and skills against those of their faculty, and stated that gender had no influence on either the interns' or faculty's ratings. However no data are presented in this brief research report to substantiate this conclusion.

Herbert et al. (1990) clearly set out to analyse the effect of gender on 142 third year obstetrics and gynaecology students' assessments of their performance against grades assigned by different groups (faculty, residents) and using different methods (clinical activities, written exams, oral examinations). The authors concluded that in terms of both departmental ratings and self-ratings for all methods of evaluation there were no differences attributable to gender (range of p values 0.07 to −0.85).

Woolliscroft et al. (1993) attempted to identify the factors that influence third year medical students' (n137) initial self-assessment of their clinical performance. Weak and negative correlations were found between self-assessments and college exam results but no statistically significant difference was found relating to gender (no p values presented).

Parker et al. (2004) looked at the ability of 311 family medicine residents to predict (i.e. self-assess) their performance on the in-training examination (ITE), regarded as an objective measure of medical knowledge. They found that residents demonstrated little ability to predict their examination scores (all Pearson correlations in 9 subject areas were less than 0.3) and there was no difference by gender.

Sommers et al. (2001) investigated how several variables including gender would affect physician faculty members' perceived self-efficacy for performing nine key professional role functions. They found that women (n21) had lower self-efficacy scores than men (n31) but that the difference was not statistically significant (p values ranged from 0.04 to 0.84 in the nine areas).

An example of contradictory evidence is found in the study by Evans et al. (2005). They examined the self-assessment skills of 50 surgeons in assessing their performance in removing a tooth. In using a checklist scale they found a significant difference between the mean scores of the assessors and male and female scores. Both males and females over-scored themselves compared to their assessors with males significantly more likely to do so than their female counterparts (difference in means (males – females) = 1.94 (95% CI = 0.26 to 3.62, $p = 0.03$)). However the same comparison with a global rating scale found no difference between males and females (difference in means (males – females) = 0.09 (95% CI = −3.36 to 3.55, $p = 0.96$)). In investigating reasons why individuals cannot assess they found no statistical difference between males and females on either of the theories they were investigating i.e. impression management (trying to convey a favourable impression) and self-deception (lack of insight). However the authors recognise that the sample sizes were too small to provide definitive evidence (32 males, 18 females).

The number of studies analysing gender differences highlights the interest in this particular aspect of self-assessment. A number of studies found no difference in the ability of males and females to self-assess. However in terms of confidence there does appear to be a trend for males to express higher levels than their female counterparts. As with most research in this area however Leopold et al. (2005) found differing evidence depending on when the confidence measurement was taken. This study typifies the inconclusive nature of evidence in the analysis of gender differences which will no doubt continue to be a fertile ground for future research.

Cultural differences

In comparison to investigations about the effects of gender (discussed here) and experience (discussed later under *Clinical Skills*), research into race and cultural differences is relatively scarce. Woolliscroft et al. (1993) correlated self-assessments and college exam results in third year medical students but found no statistically significant difference relating to race (no p values presented). Fitzgerald et al. (2003) concur that self-assessment accuracy is not

related to ethnicity from a series of studies they have undertaken.

Insight

As outlined in the *previous research* section, a series of studies on psychology students (Kruger & Dunning 1999) explored the hypothesis that incompetent students over-estimate their ability because their incompetence denies them the ability to recognize competence or lack of it, either in themselves or others. The most competent students tended to underestimate their performance, but improved their accuracy of self-assessment after benchmarking, whereas the less competent students tended to be more inaccurate after viewing others' performances. Increasing the competence of these students in logical reasoning increased the accuracy of their self-assessments, apparently by improving their metacognitive skills. Various researchers, including Hodges et al. (2001), have tested these hypotheses in clinical self-assessment settings.

Several of the higher quality papers reviewed addressed the relationship of the accuracy of self-assessment with competence, academic ability or insight into their performance.

Bryan et al. (2005) in a study of 213 first year medical students on an anatomy course stated that students with higher grades underestimated their own performance, whilst those doing poorly tended to overestimate their performance. They did not provide figures to substantiate this assertion, but did find that self rating scores were weakly positively correlated with the final grades ($r = 0.14$, $p = 0.04$).

Edwards et al. (2003) asked third year students on an obstetrics and gynaecology clerkship to estimate their final examination and clerkship grades at the beginning of the clerkship, and again just prior to the final examination. Complete sets of grades and predictions were obtained from 1139 students out of 1152. Students were more likely to accurately predict their clerkship grade than their examination grade, but for both estimates, the students ranked in the lowest third were more likely to overestimate their grades than those in the top third, who tended to underestimate their grades. The logistic regression results with 'overestimate' as the modelled outcome give odds ratios of 4.38 (CI 3.79 – 5.06) for lower versus upper third of students, and 1.90 (CI 1.66 – 2.18) for middle versus upper third of students.

Parker et al. (2004), asked 311 family medicine residents to estimate their performance in nine content areas of an in-training examination. They also found that high scorers tended to underestimate their scores and low scorers to over-estimate them. The most accurate predictions were made by the students in the middle two quartiles.

Leopold et al. (2005) examined the confidence and self-assessment of performance of 93 practitioners attending an educational session on knee injection, in relation to assessment by trained observers. Their rationale was that professionals must decide whether they have the competence to undertake a procedure, and that this decision is based on their level of confidence, as well as their background, education and skill. They found an initial significant but inverse relationship between confidence and an objective measure of performance before instruction ($r = -0.253$, $p = 0.02$), that is greater confidence was associated with poorer performance. They also found that confidence before instruction was strongly and directly correlated with the participants' assessment of their own performance ($r = 0.42$, $p = 0.001$ and therefore concluded that confidence was associated with overestimation of self-assessed performance. The effect of instruction on self-assessment was also measured and this is described in the relevant section below.

In a study of 25 resident physicians (Millis et al. 2002) self-assessment scores for an interview with a standardized patient (SP) were compared with those of the standardized patients and those of faculty. There was reasonable correlation between faculty and standardised patient ratings, (r_c 0.50, 95% CI 0.16 to 0.73) but lack of correlation between standardised patient and physician self-ratings (r_c 0.11, 95% CI -0.28 to 0.47). The resident physicians who were rated poorly by the SPs tended to rate themselves as high as physicians who were highly regarded by the SPs.

Woolliscroft et al. (1993) examined the clinical self-assessments of 137 out of 142 third year medical students compared with external measures of performance including the Medical College Admission Test (MCAT) and students' college grade-point averages (GPAs). Students in the lowest quartiles for both the GPAs and MCAT scores rated themselves highest for all skills except application of knowledge, for which students in the top quartile had a higher mean.

Mandel et al. (2005) compared the self-assessments of 74 out of 92 surgical residents with faculty ratings on two assessment measures, open surgical skills and an external global skills checklist. There was a high correlation between residents and faculty ratings on specific tasks and global skills. Unlike other studies in this section, these authors did not find that residents with poor skills were unaware of their deficiencies.

The literature reviewed contains several instances of over-estimation by poor performers, and under-estimation by those who perform well. These studies reinforce the ideas of Kruger and Dunning who argued that those who lack competence also lack the meta-cognitive skills to recognize their poor performance. Dunning (2006) explores this idea in more depth in a recent paper, suggesting that "people misjudge their incompetence not because of a lack of honesty with themselves, but rather because of a lack of the essential cognitive tools needed to provide correct self-judgments". An alternative explanation might be that such results merely reflect poor correlations between self-ratings and faculty or other assessments. Hence, rather than drawing on a psychological defence mechanism to account for the discrepancy between different raters, this finding could indicate a central tendency or regression to the mean in self-assessments. It is interesting, however, that in the Mandel et al. (2005) study it was in the area of practical skills in which the poorer performers' estimates correlated with faculty ratings and with higher scorers' estimates. This will be discussed further in the section on practical versus cognitive skills.

External factors

The purpose of the self-assessment task

In our reading of the literature it became clear that authors seldom gave information on whether or not participant self-assessment contributed to the final marks of the student or if the student self-assessment was seen by the tutor/external assessors prior to their mark being attributed.

This is important as in the first of these scenarios there may be pressure on the student to inflate their marks in order to improve their grades, reducing the apparent accuracy of their self-assessments. The impact of the second is more complex, some may see their self-assessment as a means of pressuring their tutor into giving a higher mark (it may be easier for a tutor to give a D to a student who self-assesses as D rather than one who self-assesses as B) while others may be too modest to suggest a high score even if they think they might achieve it.

We could find only one high quality study exploring the impact of either of these arrangements. Evans et al. (2005) explored the possible influence of self-deception as a possible reason for the discrepancy between self (surgeons') and assessors' ratings. They asked dental surgeons to rate their skill following removal of a third molar observed and rated by two assessors (who had good inter-rater reliability) and in addition the Paulhus Deception Scale 7 (PDS) (Paulhus 1998) was simultaneously administered. This is a validated 40 item questionnaire that measures an individual's tendency to give socially desirable responses on questionnaires. There are two components of this scale, Impression Management (IM) and Self-Deception Enhancement (SDE). Impression management refers to the tendency to give inflated self-descriptions by 'faking or lying' and to deliberately convey a favourable impression ('faking good') whereas self-deception enhancement indicates overconfidence and lack of insight. Seventy per cent of surgeons had impression management scores suggesting that they may have been deliberately trying to give a favourable impression. These IM scores correlated significantly ($r=0.45$, $p=001$) with the inability to assess their own surgical skills. Although 30% of the surgeons in this study showed lack of insight, that is to say they scored high or very high for self-deception enhancement, there was no evidence to suggest this affected their opinion of their surgical performance.

Further research exploring the impact of the purpose of self-assessment on its accuracy is required. Additionally research is needed to explore the impact of student self-assessment on external assessment.

Practical skills versus theoretical knowledge

Few studies have specifically set out to determine if self-assessment of cognitive skills differs from that of practical skills.

Edwards et al. (2003) compared the self-assessment skills of obstetrics students and found that a higher proportion of students were able to predict their clerkship grades (based on performance) than their grade by examination (56% v 31% at the start of the attachment and 61% v 32% at the end, both $p<0.001$). However, Fitzgerald et al. (2000) compared self-assessment of two sets of skills, which they described as cognitive (chest-pain questions, EKG analysis, x-ray analysis) and performance (examination of breast, chest-pain patient, unconscious patient, paediatric examination, communication skills). They found no difference in accuracy of self-assessment between either type of task.

Additionally there is evidence from other good quality studies which seems to show that practical tasks, particularly surgical tasks, appear to be amenable to self-assessment especially if feedback on performance is included. We found several papers which suggested that students had at least moderate skill in self-assessment of performance or practical skill.

Woods et al. (2004) surveyed 266 American physicians about their "comfort" (assessed on a 4 point scale) with differentiating between smallpox and chicken pox and tested them with a simple 4 question knowledge test and a visual diagnosis using photographs. 178 physicians responded. In logistic regression controlling for predictive variables (general experience, experience of rashes and speciality) only 'comfort' in diagnosis was predictive of knowledge of small pox diagnosis (OR 2.2, 95% CI 1.4 – 3.3). No parameter was found to be predictive of performance in identifying smallpox from photographs.

Ericson et al. (1997) found that dental students using performance guidelines in the area of cariology (1,373 diagnostic, preventative and restorative procedures) agreed with their tutors in 87% of assessments.

Ward et al. (2003) in a small study explored the self-assessment skills of 28 senior resident surgeons in laparoscopy. They demonstrated a correlation of $r=0.50$, $p<0.01$ immediately after conducting the surgical procedures which rose to $r=0.63$, $p<0.01$ after review of their videoed performance.

Similarly Mandel et al. (2005) compared self-assessment of proficiency on a variety of surgical bench procedures with the reliability-tested Objective Structured Assessment of Technical Skills (OSATS) in 74 obstetrics and gynaecology residents. They demonstrated high correlations with both open procedure skill ($r=0.74$, $p<0.001$) and laparoscopic skills ($r=0.67$, $p<0.001$).

Evans et al. (2005) showed modest agreement (intra-class correlation co-efficient of 0.51) between assessors and fifty dental surgeons completing a checklist on performance of extraction of a mandibular third molar.

Lane & Gottlieb (2004) compared fifty third year medical student self-assessments of interviewing skills using a 21-item five point self-assessment scale with two faculty members' assessments. Medical students disagreed with faculty in their assessment 14% of the time, but this reduced to 7% following feedback.

Weiss et al. (2005) examined the self-assessment skills of 47 third year medical students on an obstetrics and gynaecology rotation. Skills were examined in five areas: fund of knowledge, personal attitudes, clinical problem solving skills, written/verbal skills and technical skills. Self-assessments were correlated with exam results and faculty and resident ratings. They found a statistically significant

weak to moderate, positive correlation between students' self-assessment and final clerkship grade for written/verbal skills ($r=0.390$, $p=0.002$). A statistically significant agreement between raters was also revealed for written/verbal skills ($p=0.003$). Weak, non-statistically significant, positive relationships were revealed for fund of knowledge, clinical problem-solving and technical skills. A weak, negative, non-significant relationship was revealed for personal attitudes, and there was no statistically significant relationship between students' prediction of their exam score and categorized true score ($r=0.49$, $p=0.717$). This leads the authors to conclude that at the end of their obstetrics and gynaecology clerkship, third-year medical students are better at assessing their technical and written/verbal skills than their global fund of knowledge and personal attitudes.

Leopold et al. (2005) explored the impact of education and feedback on self-assessment of skill in the performance of a simulated knee joint injection. Ninety three practitioners were randomised to receive skills instruction through a manual, a video or hands-on instruction. Each participant performed one injection before and after instruction. All participants completed pre and post-instruction questionnaires on confidence and provided self-assessments of performances before and after instruction. Before instruction, participants' confidence was significantly inversely related to competent performance ($r=-0.253$, $p=0.02$). After instruction, performance improved significantly in all three training groups ($p<0.001$) with no significant differences in efficacy detected. After instruction, confidence correlated with objective competence in all groups ($r=0.24$, $p=0.04$); however, this correlation was weaker than the correlation between the participants' confidence and their self-assessment of performance ($r=0.72$, $p=0.001$).

In contrast to this, however, Rudy et al. (2001) showed poor correlation ($r=0.19$, NS) between self and faculty assessment in communication and interviewing skills in 97 first year medical students (although good correlation $r=0.50$, $p<0.0001$) between faculty and peer assessment of the students).

Antonelli (1997) showed relatively good correlation ($r=0.49$, $p=0.0006$) between global self-assessment of skill in second year medical students and perceptors' final grades but confidence in self-assessment skill was not correlated with accuracy of self-assessment. Students in this group, however already had received two third's of their year examination results and so were in a good position to predict their final score.

However, there were five included papers that failed to find a correlation between self and external assessment of knowledge in the areas of

- medical knowledge (self-assessment versus the In-training examination) in residents in family medicine (Parker et al. 2004),
- assessment of performance in undergraduate PBL tutorials (Sullivan et al. 1999; Reiter et al. 2002),
- general practitioner knowledge of thyroid disorders and diabetes (Tracey et al. 1997),
- general practitioner knowledge of techniques for assessing evidence based medicine (Young et al. 2002),
- residents' knowledge of critical care as assessed by MCQ (Johnson & Cujec 1998).

Fitzgerald et al. (2003) report a longitudinal study of medical students' self-assessment ability over three years. They noted this deteriorated in the third year. However, the examination format, which was OSCE based, was considerably different from traditional knowledge based exams they had previously sat and the authors posited that rather than the deterioration in self-assessment ability being due to increasing experience, it was due to the format of the examination.

It is not clear why practical skills may be better self-assessed than knowledge, but it could be that their outcomes are harder to dispute so the potential for self-deception about one's abilities is less. This may not apply, however, to interpersonal skills which seem relatively poorly self-assessed in the absence of structured feedback.

Factors influencing self-assessment

What factors can improve the development of self-assessment skills?

This section of the report focuses on studies which report that self-assessment skills can be improved. The Kruger and Dunning (1999) study, already referred to above, involved a series of psychological experiments in which they identified that people vary in their ability to self-assess. Of particular importance are the two groups who either over-rate or under-rate themselves. Those in the top quartile who under-rated their abilities were able to improve their self-assessment rating when shown the results of other people's work. This process helps the able student to benchmark their ability in relation to the ability of their peers, resulting in a more accurate self-assessment. The improvement in the accuracy of self-assessment has only been demonstrated for able students who previously under-rated their performance. Kruger and Dunning noted that students in the bottom quartile consistently overrated themselves despite any benchmark feedback. Self-assessment in this group was improved only by educational input to increase the level of knowledge. Thus level of knowledge or skills needed to be raised in order to improve the accuracy of self-assessment.

Video feedback and benchmarking

The importance of feedback as a tool to increase the accuracy of self-assessment was referred to by Gordon (1991). Ward et al. (2003) reported on whether self-assessment accuracy improved following video feedback after completing a surgical procedure and comparing it with a validated gold standard of expert raters. The 26 surgical residents rated their performance immediately after completing the surgical procedure. Their ratings were moderately correlated with the expert ratings ($r=0.50$, $p<0.01$). The correlation increased significantly after the residents viewed a video of their

performance and then repeated the self-assessment ($r = 0.63$, $\Delta r = 0.13$, $p = 0.01$). This study does suggest that viewing one's own performance and then completing a self-assessment is more accurate than merely relying on recall of one's own performance. Then the authors asked the residents to view four videos that represented a range of abilities, thus providing benchmarks for each level of skill. The authors expected that knowing what the standard looked like at each level would lead to a further improvement in the self-assessment accuracy of the resident's own level of skill. However no further improvement was identified and the authors postulated that this may be due to the senior skill level of the surgical residents who would have already had a good knowledge of the range of levels of performance. The margin for further improvement therefore in these circumstances would have been too small to detect a significant difference.

A similar study using benchmarks was conducted by Martin et al. (1998). The study involved 25 first and 25 second year family residents. The residents were observed by two experts while conducting a complex consultation with a standardized patient about suspected child abuse. The experts assessed the residents and the residents self-assessed their performance using the same scale. The residents were then asked to assess four benchmarked performances to determine whether the residents could identify the different benchmarked performances and whether they would match expert opinions. Following the benchmark tasks the residents were asked to reassess their own performance. The first self-assessment had a low correlation with the expert rating ($r = 0.38$), but the correlation with experts increased significantly ($p < 0.05$) after viewing the videos and re-assessing themselves ($r = 0.52$). The change in self-assessment after viewing benchmarked performances brought the assessments closer to the ratings used by experts, suggesting they were using the scale in a similar way. The mean resident–expert correlation on the benchmarked tapes was quite high (0.72) but there was quite a wide range (0.57 to 0.89). Further analyses found that the ability to correctly benchmark the videos was not related to either the ability to perform the task or the ability to accurately self-assess.

Video and verbal feedback

Lane and Gottlieb (2004) videoed the performance of 60 students conducting medical interviews and then asked students to self-rate their performance on a Likert scale that covered 21 key elements. The authors reported that the trend was for performance to improve from first to second time (319 of 432 instances, or 74% of the time). Also agreement between the rating of the tutor and those of the students improved on the second performance (14% down to 7% of errors) with a significant decrease in the rate of inaccurate assessments ($p = 0.001$). Feedback from the tutor and from viewing oneself perform was identified as the stimulus for the improvement in performance. The increase in agreement on the rating scale was again linked to feedback from the tutors who gave their views on how good the performance was and why, thus enabling the student to recalibrate what a good performance would look like.

Instruction

Leopold et al. (2005) conducted a before and after study with 93 practitioners who were randomly assigned to receive one of three instructions to improve skills on giving a knee injection. The three types of instruction were: printed manual, video and hands-on instruction. The practitioners completed a self-assessment before and after the interview. Before the intervention increased confidence was related to poorer performance ($r = -0.253$, $p = 0.02$). After the instruction performance improved significantly in all groups ($p < 0.001$), but there was no significant difference between groups. The correlation changed after the intervention from a negative to a positive correlation, showing that confidence was related to performance, but the correlation was weaker ($r = 0.24$, $p = 0.04$). The authors concluded that even low intensity forms of instruction improved confidence, competence and self-assessment.

Experience

There is some evidence that increased experience in a skill or knowledge is also reflected in higher scores on a self-assessment scale. Studies examined two particular aspects of experience. The first is the relative level of experience of the participants in relation to their clinical knowledge, skills or expertise, for example novice versus expert. Typically this might involve first year undergraduates being compared to third year undergraduates. The second aspect of experience explored is the effect of exposure on an individual's ability to self-assess. This involves examining proficiency before and after an intervention or experience e.g. attendance on a rotation. The objective is to determine whether exposure to a skill or experience increases an individual's accuracy in assessing their performance as they become better accustomed to the respective task or skill and acquire better knowledge.

Novice versus expert

Wilkerson et al. (2002) investigated the effects of an enhanced curriculum in cancer prevention on medical students' (n333) knowledge and self-perceived competency in the use of counselling and screening examinations during the first three years of medical school. This enabled them to compare the three different years of students with varying levels of knowledge and experience. They reported that students' knowledge of cancer prevention significantly improved over time (e.g. third year students scored significantly higher than the years below them, $p < 0.001$). The reported improvement in the self-assessed skills of counselling and screening skills was correlated to hands-on practice. When practice was removed, as in the second year, the improvement in self-assessed skills was absent. This finding suggests that hands-on practice provided an opportunity for knowledge and skills to be tested out and providing the individual with some feedback increased the self-rated competencies.

Herbert et al. (1990) evaluated the effect of previous clerkship experience on the actual grades that 142 third year students achieved on a six week obstetrics and gynaecology clerkship. There was no correlation between the grades

achieved and previous clerkship experience and more experience did not affect students' ability to self-assess. Unfortunately no data is presented to verify this conclusion.

Sommers et al. (2001) specifically examined the length of faculty members' (n54) experience on their self-perceived efficacy for carrying out key medical functions. They concluded that time in faculty did not have any significant effect on the total self-efficacy scores for the nine professional role functions examined i.e. increasing the length of time in a faculty position did not influence self-efficacy scores (p values ranged from 0.042 to 0.78 in the nine areas). Furthermore they found no statistically significant association between age and the total self-efficacy score or that for the nine individual areas investigated (no data are presented to verify this finding).

Leopold et al. (2005), summarized above, also reported that prior to the intervention, practitioners with more expertise rated themselves higher than their peers, although their performance was not significantly better. After the intervention there was again no correlation with experience and greater performance (as measured by increased years in practice or by giving three or more injections).

Paradise et al. (1997) asked 206 physicians who rated their skills as above average in evaluating cases of suspected sexual abuse to examine seven simulated cases by means of a questionnaire. The physicians' descriptions and interpretations of the simulations were compared with consensus standards developed by an expert panel. In three of the simulations the most experienced physicians resembled the panel more closely than did the less experienced ($p \leq 0.001$). This leads to the conclusion that among physicians who self-rate themselves as skilled, assessments made by more experienced physicians may relate more closely to consensus standards than those made by less experienced physicians.

Exposure and feedback

Edwards et al. (2003) conducted a before and after study involving 1,152 students comparing the differences between predicted and actual final examination and clerkship grades. This was an extensive study over ten years of third year students ($n = 1{,}152$) in an obstetrics and gynaecology clerkship. Students were more likely to correctly predict their clerkship grade than their examination result, at the beginning (56% vs 31%, $p < 0.001$) and at the end (61% vs 32%, $p < 0.001$). The authors reported that students who had slightly shortened placements (6 weeks compared with 8) were 3.6 times more likely to overestimate their clerkship performance than the students on the 8 week placement. Also students who did the clerkship earlier on in their careers (during the autumn semester) were 1.55 times more likely to overestimate their performance than those who did it later on in the spring semester. The authors suggest that on-going feedback during the clerkship may have had an effect on the greater predicted accuracy of the clerkship grade compared to the exam grades. The authors postulate the importance of feedback, which they suggest plays a mediating role in accurate self-assessment.

Zonia and Stommel (2000) evaluated the difference between interns' self-assessments (n73) and those made by their faculty. In terms of experience they found that interns' self-ratings and equivalent faculty ratings consistently increased in the first five months of their rotations ($p = 0.001$). However after the fifth month the ratings reached a plateau.

Gruppen et al. (2000) ran a study which aimed to correlate how amounts of study time linked to changes in self-assessed diagnostic capabilities over the course of a three month clerkship. The subjects were 107 medical students in three consecutive cohorts of an internal medicine clerkship. This was a before and after study which correlated a self-assessed measure of confidence at the start and finish of the clerkship with an estimate of time spent studying respective topics. The researchers found a modest but positive correlation (mean co-efficient = 0.25, SD = 0.20; 95% CI 0.21 to 0.29) leading them to conclude that spending more time on a given topic resulted in an increase in self-assessed diagnostic skill for that subject. They cautioned that individual variation influenced the strength of the relationship, it being much stronger for some students than others (range = −0.23 to 0.89).

Eva et al. (2004) in a study of 265 Canadian medical students found no evidence that performance in self-assessment improved over 2.5 years of schooling. They did find that students who estimated their examination performance after sitting the examination were more accurate than those who predicted their score before taking the examination.

The level of experience of those self-assessing raises an interesting question in the literature, namely whether it is experience in the knowledge or skill being assessed that determines self-assessment ability or experience of self-assessment itself which is most important in determining accuracy. Ward et al. (2003) examined the self-assessment accuracy of 26 surgical residents and whether self-observation of their performance by video and the opportunity to view benchmark videos of performance would improve their self-assessment ability. Initially there was a moderate correlation between experts' evaluations and residents' self-evaluations ($r = 0.50$, $p < 0.01$). They found that self-observation did improve self-assessment ability ($r = 0.63$, $\Delta r = 0.13$, $p < 0.01$) but exposure to benchmarked performances did not ($r = 0.66$, $\Delta r = 0.03$, NS). This leads them to conclude that ability to self-assess is related in this case to surgical experience rather then self-assessment experience.

In summary, these studies highlight the importance of both feedback on performance, and of increasing knowledge of the task to increase understanding and recalibration of what a good performance involves.

Perceptions and attitudes towards self-assessment

We set out to determine the attitudes towards and perceptions of learners and teachers to self-assessment. However, few papers in our review made more than a passing reference to this feature of self-assessment and, among those that did, no single paper met our quality threshold for inclusion. There were no studies that focused on perceptions alone; these were always of secondary consideration.

Whilst the evidence is not robust, the papers we examined would seem to suggest a favourable response towards self-assessment activities on the whole by participants. There is occasional indication of stressful and threatening reactions experienced by students in some studies but this requires further exploration.

The acceptability of self-assessment as an educational tool is assumed rather than explored in the literature. There is an urgent need for high quality research in this area. The lack of a robust evidence-base about attitudes towards self-assessed activities is somewhat contrary to their importance in practice for identifying leaning needs and maintaining competence in health professional behaviour. The dearth of robust qualitative research is of particular concern in this field.

Discussion

The research questions addressed by this review sought evidence for the effectiveness of self-assessment interventions to

- improve the accuracy of learner perception of their learning needs,
- promote an appropriate change in learner learning activity,
- improve clinical practice,
- improve patient outcomes.

Subsidiary research questions addressed factors affecting the accuracy of self-assessment, and learners' and teachers' perceptions of and attitudes towards self-assessment.

Overall, it appears that the review, despite a robust methodology, was largely unable to answer the specific research questions, and provide a solid evidence base for effective self-assessment. No papers were found which satisfied Kirkpatrick's hierarchy above level 2, and we found no studies which looked at the association between self-assessment and resulting changes in either clinical practice or patient outcomes.

However, in terms of our subsidiary questions, while no indisputable evidence was found, our review did identify several factors which appear to influence self-assessment. In order to increase our understanding of the conditions which are associated with accurate self-assessment, it is recommended that these areas would merit further research.

Positive findings

An interesting conclusion across a number of studies was that individuals are far more able to accurately assess their peers' ability than their own. Peer assessments also appear to be more in line with faculty assessments of performance than self-assessments. This could be important when considering methods of validating self-assessment.

Ability and experience would appear to have some impact on self-assessment, with several papers exploring the relationship between accuracy of self-assessment and competence or academic ability. The findings from these studies broadly support the idea that competent practitioners are reasonably accurate in their self-assessment, and it may be possible to improve this accuracy. On the other hand, people who lack competence are less likely to be aware of their deficiencies as evidenced by self-assessment, and to be less responsive to strategies for improving accuracy. This has important implications, and is worthy of further research.

There is some evidence from our review that practical skills may be better self-assessed than knowledge. As noted in the results section, this could perhaps be explained by the fact that the outcomes of practical skills are harder to dispute and so the potential for self deception about one's own abilities is less. Observable performance also lends the opportunity for direct feedback.

The importance of feedback and benchmarking has been identified in a small number of studies in our review as increasing the accuracy of self-assessment by increasing the learner's awareness of the standard to be achieved.

Inconclusive or negative findings

Gender is an obvious starting point in looking for potential reasons for differences in self-assessment outcomes. Although there were more papers examining differences by gender than any other type of sub-analyses, most of the evidence here was inconclusive or contradictory and may have been relative to the type of activity under consideration.

There was no high quality evidence to suggest that race or culture impact on an individual's ability to rate themselves objectively.

In the context of how self-assessment is perceived by learners and teachers, our review suggests that the acceptability of self-assessment is seldom explored. Of those which did address this, there would seem to be a favourable response to self-assessment activities by participants, although self-assessment may be stressful for some students and even potentially threatening. Attitudes towards self-assessment may be influenced by the purpose of the self-assessment activity itself, that is whether self-assessment is undertaken for formative or summative outcomes. The need for high quality research is particularly urgent in this field.

Strengths of our review

At the start of the project, considerable time was spent developing a rigorous methodology with which to conduct the review. Agreeing an explicit definition of self-assessment was itself a complex activity and this will be addressed later.

As noted in the Methods section, we developed a rigorous review process, which incorporated several iterative stages.

- Development and use of a standardized coding and quality checklist adapted form published tools (http://www.phru.nhs.uk/casp/casp.htm)
- All papers were reviewed independently by pairs, with recourse to an adjudicator to resolve disagreements
- Iterative process of reviewing and discussing papers and if necessary revisiting the full text
- Regular discussion between pairs and with the whole group to clarify concepts

- Peer review/feedback from presentations at international conferences (ASME, AMEE and Ottawa).

Difficulties encountered

Some 'teething problems' were experienced, perhaps inevitably, around the development phase of the electronic coding form. Overcoming these has benefited a subsequent review which is using a similar e-form.

Although a large number of papers resulted from our original search ($n = 5{,}798$), only a small proportion were of sufficient academic rigour to be included in our review ($n = 32$). Research on self-assessment has been fraught with methodological problems, and this is reinforced by our review where reasons for exclusion included no clear definition of self-assessment, inadequate information on sampling strategies, and insufficient reporting of methods and analysis. Similar concerns about the quality of published research in self-assessment have been expressed by Davis et al. (2006). These authors conducted a more focused review, limited to a comparison of physician self-assessment with observed measures of competence. Despite this more specific context, only 17 out of 725 papers met all the inclusion criteria. One of the implications from both reviews is that the peer review process in many journals may need to be more rigorously implemented.

Most of the papers of sufficient quality to be included in our review concentrated on judging the accuracy of self-assessment by comparison with some external standard (as was the focus of the Davis review), but as outlined above there are problems with this approach. This left few papers selected for our review that actually addressed our specific research questions.

Self-assessment, no matter how it is defined, is a complex concept which does not lend itself to objective measurement. It may be, therefore, that our conclusions were limited by our definition of self-assessment, and that the outcome of our review would have been more definitive if we had used a broader definition, particularly one which takes account of meta-cognitive skills. Despite attempts to standardise our approach to inclusion and exclusion of papers, there is inevitably a subjective element to making this final judgement, and this may have resulted in some borderline papers being excluded.

Philosophy of self-assessment and problems of definition

In our definition we said that self-assessment is "a personal evaluation of one's professional attributes and abilities against perceived norms".

Very few of the papers that we reviewed defined the concept of self-assessment that they were researching. The majority of them set out to determine the 'accuracy' of self-assessment in terms of quantitative comparisons with external measures or 'expert' ratings. Ward et al. (2002) point out the problems with these types of studies, namely lack of validity and reliability of the 'gold standard', the likelihood of differential use of scales among students, and problems of group level analyses.

Colliver et al. (2005) concur with Ward et al. (2002), and go further in suggesting that this type of quantitative analysis of 'guess your grade' type studies is not relevant to the daily ongoing self-assessment of practice. The latter involves the recognition of specific deficits in knowledge or skills in the context of the clinician's practice. They make the point that self-assessment for ongoing self-directed learning is a qualitative exercise, concerned with specific subjects in an individual context. This would lend itself to a narrative approach about an individual's clinical knowledge and skill, and indeed could not be quantified. They suggest that this personalized assessment in practice should be the target of research, and that this is beyond the conventional quantitative research paradigm.

Eva and Regehr (2005) follow a similar thread when they argue that although simple definitions of self-assessment are attractive, they tend to cause difficulties because they do not allow for the complexity of the concept. They suggest the adoption of a different paradigm, in which professionals constantly self-assess in terms of their own strengths and weaknesses in relation to situations that they experience. The ability to identify one's weaknesses can lead to knowing when to ask for help with a case, or to setting appropriate learning goals. Being aware of one's strengths allows one to persevere with a correct course of action despite initial setbacks, and to set realistic, challenging, but achievable learning goals.

The authors point out that self-concept, "a relatively sweeping cognitive appraisal of oneself", and self-efficacy, "a context-specific assessment of competence to perform a specific task"' will both influence self-assessments. They argue that self-efficacy differs from self-assessment in that it influences our performance, a strong sense of self-efficacy leading to a greater chance of success.

In our introduction, some reference was made to how we defined self-assessment for our review, and the difficulty this raises in the context of self-referent thinking. Wooliscroft et al. (1993) draw on psychological literature to argue that an individual's view of self, or 'self-concept' results from external feedback and introspection. Accurate self-assessment clearly depends on congruence between self-representation and reality, but these authors argue that over time, self-representation becomes increasingly resistant to change despite feedback. This reinforces Gordon's (1991) finding that self-assessment did not always change as a result of external evaluative information. It is not clear however why low achievers are more likely than high achievers to overestimate their abilities, although some authors suggest some kind of psychological 'defence' mechanism (Woolliscroft et al. 1993). Such psychological self-protection strategies could also explain the studies that found that generally we assess others more accurately than we assess ourselves.

In the psychological literature, the concept of self-efficacy originates from a theoretical basis which emphasises the importance of feedback in shaping subsequent action (Bandura 1977, 1986). Like Woolliscroft's explanation of self-representation, self-efficacy thus incorporates

environmental (external) and cognitive (internal) factors on learning behaviour. Eva and Regehr (2005) have defined self-efficacy as "an individual's judgement of her capabilities to complete a given goal" (p. 548). These authors argue that the literature on self-assessment focuses on 'accuracy' (reinforced by our review) while research around self-efficacy focuses on the consequences of particular self-efficacy beliefs and their impact on future performance of tasks, which is arguably a key outcome. They also address the need to consider a third source of variation in self-assessment capacity, namely the meta-cognitive factors which affect individual judgements about learning, and in particular how individuals process the feedback and judgements about their performance made by others. As already noted, Kruger and Dunning (1999) hypothesised that deficient self-assessment may result from lack of meta-cognitive skills, and cited some evidence that improving meta-cognitive skills (in this case logical reasoning) improved self-assessment accuracy. Eva and Regehr (2005) have reviewed the research paradigms of several different but related disciplines. They express the view that the literature on reflective practice supports the idea of moving away from the concept of self-assessment as a 'conscious meta-cognitive and usually post-hoc summative process', and that safety in professional work requires that self-assessment be conceptualised as an ongoing 'reflection-in-action', constantly monitoring one's ability to deal with the emerging situation.

In a paper published since we commenced our review, Dunning (2006) argued that the flawed nature of self-assessment could result from individual cost/benefits analysis – a theory well-documented in the context of risk-taking health behaviours. Strategies suggested for correcting mistaken self-judgements include recognising the importance of listening to external feedback, especially from peers, or improving meta-cognitive skills to be more realistic in the light of external 'evidence'. The third strategy proposed by Dunning is simply to adopt 'cognitive repairs'–in other words recognise that self-assessment is often inaccurate, and make appropriate allowances.

The accuracy of self-assessment as a measure of clinical performance may in fact be no worse (and no better) than any other single judgement of competence. There is a large body of evidence to suggest that many judgements (and methods) are required before stable and reproducible ratings of performance can be obtained (Carline et al. 1989; van der Vleuten & Swanson 1990; Williams et al. 2003). Perhaps the nature of the self-assessment task is the issue here. In setting appropriate goals for learning, individuals must be able to identify their own weaknesses as well as their own strengths in the context of good professional practice. Relying solely on a self-assessment tool may be insufficient to determine the full extent of learning needs. In a paper already referred to earlier in this review, Amery and Lapwood (2004) found a clear disparity between respondents' self-rated skills and their educational requirements as derived from personal diaries. The gap between perceived and actual need led these authors to make a case for multiple assessment tools to fully identify the ongoing training required by health professionals. In this study, the use of self-assessment as a single measure failed to pick up unmet educational, training and support needs in areas of clinical practice. To discount self-assessment as wholly inaccurate or flawed, however, is rather to miss the point. We should be aware of the limitations of self-assessment but use it alongside other sources of information to provide broader, more holistic assessments of competence and learning activity for health professionals in practice.

Future research

From the discussion above and the findings of our review, we would suggest a move away from quantitative comparison studies of the 'accuracy' of self-assessment. As Eva and Regehr (2005) point out, the problem with this paradigm runs deeper than flawed methodology of studies. They suggest that the problem is one of "a failure to effectively conceptualize the nature of self-assessment in the daily practice of healthcare professionals, and a failure to properly explicate the role of self-assessment in a self-regulating profession". Future researchers would do well to consider the relevant literatures summarised in their article (Eva and Regehr 2005) before attempting to articulate their own research questions.

Future research could shift the focus to individual cognitions about their own developing clinical competence. This might, for example, explore the kinds of cognitive pathways that underpin self-assessment and performance, to clarify the relationships between self-efficacy, self-concept, motivation, self-assessment, and performance (perceived and externally measured). Qualitative research on the influences on the judgements that people make about themselves, the effect of interactions with and feedback from peers on self-assessment, and the triggers in everyday practice that highlight learning needs would provide a platform of information on which to build. Where there is doubt about the effectiveness of self-assessment interventions, randomized controlled trials could then be constructed on a well-defined theoretical basis, to determine their effect on the accuracy of determination of learning needs, or on subsequent learning activity and change in clinical practice. Current appraisal systems and the increasing use of multi-source feedback in the health professions lend themselves to research of this nature, and could be usefully informed by such research.

Conclusion

Self-assessment is integral to lifelong learning in the health care professions. However there is evidence that in some contexts and tasks self-assessment is inaccurate. More worryingly there is evidence that those who are least able are also least able to self-assess accurately. If self-assessment is to remain the cornerstone of continuing professional development and in determining how regulatory appraisal requirements are to be met, we need to have a greater understanding of what forms of self-assessment are useful in determining learning needs, and what impact these have on future learning activities.

Our systematic review has been unable to answer these questions, but it has added weight to the arguments to consider different research paradigms to significantly increase our understanding of how self-assessment works or can be improved. We did however find themes in the literature

around self-assessment that offer clear possibilities for future research to increase our understanding of the process.

Contribution of reviewers

Brian McKinstry led the review. All reviewers conceived and designed the review. All reviewers evaluated the abstracts and relevant full text papers. All reviewers abstracted data, analysed the findings and wrote the report.

Sources of support

LTSN (now part of the Higher Education Academy), NHS Education for Scotland, University of Newcastle, University of Leeds, University of Edinburgh. Brian McKinstry is supported by the Chief Scientist Office of the Scottish Executive.

Conflicts of interest

None known

Notes on contributors

IAIN COLTHART is Research & Information Officer at NHS Education for Scotland, Edinburgh, UK.

GELLISSE BAGNALL is Educational Development Manager, NHS Education for Scotland, Glasgow, UK.

ALISON EVANS is Associate Director in Postgraduate General Practice Education (Yorkshire), NHS PG Medical & Dental Education, University of Leeds, UK.

HELEN ALLBUTT is Research and Information Officer, NHS Education for Scotland, UK.

ALEX HAIG is e-Portfolio Projects Manager, NHS Education for Scotland, Edinburgh, UK.

JAN ILLING is a Senior Research Associate at the Postgraduate Institute for Medicine and Dentistry at the University of Newcastle upon Tyne, UK.

BRIAN MCKINSTRY is Senior Research Fellow in the General Practice section of the Division of Community Health Sciences, The University of Edinburgh, UK.

Acknowledgements

Rachel Adams, Heather Peacock and Susan Hrisos for their help in the early stages of the project. Marilyn Hammick for her advice and training. Neil McManus for creating the on-line coding form.

Bibliography of Reviewed Papers

Citations that were strong enough to be informative (Table 1: www.bemecollaboration.org)

Antonelli MA. 1997. Accuracy of second-year medical students' self-assessment of clinical skills. Acad Med 72:S63–65.

Bryan RE, Krych AJ, Carmichael SW, Viggiano TR, Pawlina W. 2005. Assessing professionalism in early medical education: experience with peer evaluation and self-evaluation in the gross anatomy course. Ann Acad Med Singapore 34:486–491.

Edwards R, Kellner K, Sistrom C, Magyari E. 2003. Medical student self-assessment of performance on an OG clerkship. Am J Obstet Gynecol 188:1078–1082.

Ericson C, Christersson C, Manogue M, Rohlin M. 1997. Clinical Guidelines and Self Assessment in Dental Education. Eur J Dental Educ 1:123–128.

Eva KW, Cunnington JPW, Reiter HI, Keane DR, Norman GR. 2004. How Can I Know What I Don't Know? Poor Self Assessment in a Well-Defined Domain. Adv Health Sci Educ 9:211–224.

Evans AW, Leeson RM, Newton John TR, Petrie A. 2005. The influence of self-deception and impression management upon self-assessment in oral surgery. Br Dental J 198:765–769.

Fitzgerald J, White C, Gruppen L. 2003. A longitudinal study of self-assessment accuracy. Med Educ 37:645–649.

Fitzgerald JT, Gruppen LD, White CB. 2000. The Influence of Task Formats on the Accuracy of Medical Students' Self-assessments. Acad Med 75:737–741.

Gruppen L, White C, Fitzgerald T, Grum C, Woolliscroft J. 2000. Medical Students Self-assessments and their Allocations of Learning Time. Acad Med 75:374–379.

Herbert WN, Mcgaghie WC, Droegemeuller W, Riddle MH, Maxwell KL. 1990. Student evaluation in obstetrics and gynecology: self vs departmental assessment. Obstet Gynecol 76:458–461.

Hodges B, Regehr G, Martin D. 2001. Knowing what we know - difficulties in recognising ones own incompetence: novice physicians who are unskilled and unaware of it. Acad Med 76:S87–89.

Johnson D, Cujec B. 1998. Comparison of self, nurse and physician assessment of residents rotating through an ICU. Crit Care Med 76:1811–1816.

Lane JL, Gottlieb RP. 2004. Improving the interviewing and self-assessment skills of medical students: is it time to readopt videotaping as an educational tool? Ambulatory Paediatrics 4:244–248.

Leopold SS, Morgan HD, Kadel NJ, Gardner GC, Schaad DC, Wolf FM. 2005. Impact of educational intervention on confidence and competence in the performance of a simple surgical task. J Bone Joint Surgery 87A:1031–1037.

Mandel LS, Goff BA, Lentz GM. 2005. Self-assessment of resident surgical skills: Is it feasible? Am J Obstet Gynecol 193:1817–1822.

Martin D, Regehr G, Hodges B, Mcnaughton N. 1998. Using videotaped benchmarks to improve the self-assessment ability of family practice residents. Acad Med 73:1201–1206.

Millis SR, Jain SS, Eyles M, Tulsky D, Nadler SF, Foye PM, Elovic E, Delisa JA. 2002. Assessing Physicians' Interpersonal Skills: do patients and physicians see eye to eye? Am J Phys Med Rehab 81:946–951.

Minter RM, Gruppen LD, Napolitano KS, Gauger PG. 2005. Gender differences in the self-assessment of surgical residents. Am J Surgery 189:647–650.

Paradise JE, Finkel MA, Beiser AS, Berenson AB, Greenberg DB, Winter MR. 1997. Assessments of girls' genital findings and the likellihood of sexual abuse: agreement among physicians self-rated as skilled. Arch Pediat Adolescent Med 151:883–891.

Parker RW, Alford C, Passmore C. 2004. Can Family Medicine Residents Predict Their Performance on the In-Training Examination? Fam Med 36:705–709.

Reiter H, Eva K, Hatala R, Norman G. 2002. Self and Peer Assessment in Tutorials: Application of a Relative-ranking Model. Acad Med 77:1134–1139.

Rudy DW, Fejfar MC, Griffith CH, Wilson JF. 2001. Self and peer assessment in a first year communication and interviewing course. Eva Health Professions 24:436–445.

Sommers PS, Muller JH, Ozer EM, Chu PW. 2001. Perceived self-efficacy for performing key physician faculty functions - a baseline assessment of participants in a one-year faculty development program. Acad Med 76:S71–73.

Sullivan ME, Hitchcock MA, Dunnington GL. 1999. Peer and Self Assessment During Problem Based Tutorials. Am J Surg 177:266–269.

Tracey J, Arroll B, Barham P, Richmond D. 1997. The validity of general practitioners self assessment of knowledge: cross sectional study. Br Med J 315:1426–1428.

Ward M, Macrae H, Schlachta C, Mamazza J, Poulin E, Reznick R, Regehr G. 2003. Resident self-assessment of operative performance. Am J Surg 185:521–524.

Weiss PM, Koller CA, Hess LW, Wasser T. 2005. How do medical student self-assessments compare with their final clerkship grades? Med Teach 27:445–449.

Wilkerson L, Lee M, Hodgson CS. 2002. Evaluating curricular effects on medical students knowledge and self-perceived skills in cancer prevention. Acad Med 77:S51–53.

Woods R, Mccarthy T, Barry MA, Mahon B. 2004. Diagnosing Smallpox: Would you know it if you saw it? Biosecurity and Bioterrorism: Biodefense Strat, Pract Sci 2:157–163.

Woolliscroft JO, Tenhaken J, Smith J, Calhoun JG. 1993. Medical students' clinical self-assessments: comparisons with external measures of performance and the students self-assessments of overall performance and effort. Acad Med 68:285–294.

Young JM, Glasziou P, Ward JE. 2002. General practitioners' self-ratings of skill in evidence based medicine: validation study. Br Med J 324:950–951.

Zonia SL, Stommel M. 2000. Interns Self-evaluations Compared with Their Faculty's Evaluations. Acad Med 75:742.

Citations that were not strong enough to be informative

Adriaansen MJM, Van Achterberg T. 2004. A test instrument for palliative care. Int J Nurs Studies 41:107–117.

Alnasir F, Grant N. 1999. Student self-assessment in a community-based clinical clerkship in family medicine: a preliminary report. Educ Health 12:161–166.

Bakken LL, Sheridan J, Carnes M. 2003. Gender differences among physician-scientists in self-assessed abilities to perform clinical research. Acad Med 78:1281–1286.

Barnsley L, Lyon PM, Ralston SJ, Hibbert EJ, Cunningham I, Gordon FC, Field MJ. 2004. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. Med Educ 38:358–367.

Biernat K, Simpson D, Duthie E, Bragg D, London R. 2003. Primary Care Residents Self-assessment Skills in Dementia. Adv Health Sci Educ 8:105–110.

Chur-Hansen A. 2000. Medical students' essay-writing skills: criteria-based self-and tutor-evaluation and the role of language background. Med Educ 34:194–198.

Claridge JA, Calland JF, Chandrasekhara V. 2003. Comparing resident measurements to attending surgeon self-perceptions. Am J Surg 185:323–327.

Coutts L, Rogers J. 1999. Student Assessment and Standardized Patients - Will the Questions Never End? Acad Med 74:S128–130.

Das M, Mpofu D, Dunn E, Lanphear JH. 1998. Self and tutor evaluations in problem-based learning tutorials: is there a relationship? Med Educ 32:411–418.

Davis JD. 2002. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. Am Coll Obstet Gynecol 99:647–651.

Elliott N, Higgins A. 2005. Self and peer assessment - does it make a difference to student group work. Nurse Educ Pract 5:40–48.

Evans A, Mckenna C, Oliver M. 2001. Towards better understanding of self-assessment in oral and maxillofacial surgery. Med Educ 35:1077.

Farnhill D, Hays SC, Todisco J. 1997. Interviewing skills: self-evaluation by medical students. Med Educ 31:122–127.

Fincher R, Lewis L. 1994. Learning, experience, and self-assessment of competence of third-year medical students in performing bedside procedures. Acad Med 69:291–295.

Fincher RM, Lewis L, Kuske T. 1993. Relationships of Interns Performances to their Self-assessments of their Preparedness for Internship and to their Academic Performances in Medical School. Acad Med 68:S47–S50.

Fox RA, Ingham Clark CL, Scotland AD, Dacre JE. 2000. A study of pre-registration house officers' clinical skills. Med Educ 34:1007–1012.

Frye AW, Richards BF, Bradley EW, Philp JR. 1991. The consistency of students self-assessments in short essay subject matter examinations. Med Educ 25:310–316.

Harrington JP, Murnaghan J, Regehr G. 1997. Applying a relative ranking model to the self -assessment of extended performances. Adv Health Sci Educ 2:pp. 17–25.

Hartman SL, Nelson MS. 1992. What we say and What we do: Self-reported teaching behavior versus performances in written simulations among medical school faculty. Acad Med 67:523–527.

Hoppe RB, Farquhar LJ, Henry R, Stoffelmayr B. 1990. Residents' Attitudes towards and Skills in Counseling: using Undetected Standardised Patients. J Gen Intern Med 5:415–420.

Kaiser S, Bauer J. 1995. Checklist Self-Evaluation in a Standardized Patient Exercise. The Am J Surg 169:418–420.

Levinson W, Gordon G, Skeff K. 1990. Retrospective versus actual pre-course self-assessments. Eval Health Prof 13:445–452.

Lind DS, Rekkas S, Bui V, Lam T, Beirle E, Copeland E. 2002. Competency-Based Student Self-Assessment on a Surgery Rotation. J Sur Res 105:31–34.

Lorenz R, Gregory RP, Davis DL. 2000. Utility of a brief self-efficacy scale in clinical training program evaluation. Eval Health Prof 23:pp. 182–193.

Macdonald J, Williams RG, Rogers DA. 2003. Self-assessment in simulation-based surgical skills training. The Am J Surg 185:319–322.

Mattheos N, Nattestad A, Christersson C, Jansson H, Attstrom R. 2004. The effects of an interactive software application on the self-assessment ability of dental students. Eur J Dent Educ 8:97–104.

Miller PJ. 1999. The agreement of peer assessment and self-assessment of learning processes in problem-based learning. J Phys Ther Educ 13:26–30.

Murdoch-Eaton D. 2002. Reflective Practice Skills in Undergraduates. Acad Med 77:734.

Murdock JE, Neafsey PJ. 1995. Self-efficacy measurements: an approach for predicting practice outcomes in continuing education? J Cont Edu Nurs 26:158–165.

Pierre RB, Wierenga A, Barton M, Thame K, Branday JM, Christie CD. 2005. Student self-assessment in a paediatric objective structured clinical examination. West Indian Med J 54:144–148.

Rees C, Shepherd M. 2005. Students' and assessors' attitudes towards students' self-assessment of their personal and professional behaviours. Med Educ 39:30–39.

Sklar DP, Tandberg D. 1995. The value of self-estimated scholastic standing in residency selection. J Emerg Med 13:683–685.

Stewart J, Ohalloran C, Barton JR, Singleton SJ, Harrigan P, Spencer J. 2000. Clarifying the concepts of confidence and competence to produce appropriate self-evaluation measurement scales. Med Educ 34:903–909.

Tousignant M, Desmarchais JE. 2002. Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study. Adv Health Sci Educ 7:19–27.

Wagner C, Abholz HH. 2004. Der Effekt einer palliativmedizinischen Fortbildungsreihe auf den Kenntnisstand von Hausrzten und deren Selbsteinschtzung [Effects of CME in Palliative Medicine on GP's knowledge and self-evaluation]. Zeit Allgemeinmedizin 80:150–152.

Weerakoon PK, Fernando DN. 1991. Self-evaluation of skills as a method of assessing learning needs for continuing education. Med Teach 13:103.

Wetherell J, Mullins G, Hirsch R. 1999. Self-assessment in a problem-based learning curriculum in dentistry. Eur J Dental Educ 3:97–105.

Windish DM, Knight AM, Wright SM. 2004. Clinician Teachers Self-assessment versus learners' perceptions. J Gen Inter Med 19:554–557.

Zijlstra-Shaw S, Kropmans TJ, Tams J. 2005. Assessment of professional behaviour - a comparison of self-assessment by first year dental students and assessment by staff. Br Dent J 198:165–171.

Additional references

American Medical Association. 1992. 'Mirror, mirror: medicine enters the era of self-assessment'. ALP Observer 12:17.

Amery J, Lapwood S. 2004. A study into the educational needs of children's hospice doctors: a descriptive quantitative and qualitative survey. Palliative Med 18:727–733.

Bandura A. 1977. Self-efficacy: toward a unifying theory of behavioural change. Psychol Rev 84:191–215.

Bandura A. 1982. Self-efficacy mechanism in human agency. Am Psychol 37:122–147.

Bandura A. 1994. Self-efficacy. In Ramachaudran VS. (Ed) Encyclopedia of human behavior 4:71–81. New York: Academic Press. (reprinted in Friedman H. Encyclopedia of mental health. San Diego: Academic Press, 1998).

Bandura A. 1986. *Social Foundations of Thought and Action: A Social Cognitive Theory* (Englewood Cliffs, NJ, Prentice-Hall).

Boud D. 1995. *Enhancing Learning through Self Assessment* (London, KoganPage).

British Medical Association. 2003. *Appraisal, A Guide for Medical Practitioners* (London, BMA Publications).

Carline JD, Weinrich MD, Ramsey PG. 1989. Characteristics of ratings of physician competence by professional associates. Eval Health Prof 12:409–423.

Dornan T, Littlewood S, Margolis SA, Scherpbier A, Spender J, Ypinazar V. 2006. How can experience in clinical and community settings contribute to early medical education? A BEME systematic review. Med Teach 28:3–18.

French Medicine Association cited in: Bruneau, C., Lachenaye-Llanas, C. 2002. The French Accreditation System. Clin Gov Bull 3:8–10.

General Medical Council. 2002. *'Tomorrow's Doctors: Recommendation on Undergraduate Medical Education'* , 2nd edn (London, GMC).

Gordon MJ. 1991. A Review of the Validity and Accuracy of Self-assessments in Health Professions Training. Acad Med 66:762–769.

Gordon MJ. 1992. Self-assessment Programs and Their Implications for Health Professions Training. Acad Med 67:672–679.

Haig A, Dozier M. 2003. BEME Guide No.3: systematic searching for. evidence in medical education - Part 1: sources of information. Med Teach 25:352–363.

Hammick M, Freeth D, Koppel I, Reeves S, Barr H. 2007. A best evidence systematic review of interprofessional education. Med Teach 29:735–751.

Jamtvedt G, Young JM, Kristoffersen DT, O'brien MA, Oxman AD. 2006. Audit and feedback: effects on professional practice and health care outcomes. Cochrane Database of Systematic Reviews, Art. No: CD000259. DOI: 10.1002/14651858.CD000259.pub2.

Kirkpatrick DL. 1967. Evaluation of training, In: Bittel L. Editors *Training and Development Handbook*, (New York, McGraw-Hill), pp. 87–112.

Paulhus D. L. 1998. *Paulhus Deception Scales (PDS): The Balanced Inventory of Desirable Responding – 7. User's manual* (New York, Multi-Health Systems Inc.).

Schwarzer, R 'General perceived Self-Efficacy in 14 cultures', www.user page.fu-berlin.de/~health.world14.htm (accessed 30/06/05).

United Kingdom Central Council For Nursing, Midwifery and Health Visiting. 1999. *Fitness for Practice: The UKCC Commission Nursing and Midwifery Education* (London, UKCC).

Van Der Vleuten CPM, Swanson DB. 1990. Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med 2:58–76.

Williams RG, Klamen DA, McGaghie WC. 2003. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med 15:270–292.

Reference papers

These references did not meet the inclusion criteria but were identified through the original and follow-up search methods. They informed our thinking during the review and provided evidence of previous research on self-assessment.

Colliver JA, Verhulst SJ, Barrows HS. 2005. Self-assessment in medical practice: a further concern about the conventional research paradigm. Teach Learn Med 17:200–201.

Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. 2006. Accuracy of physician self-assessment compared with observed measure of competence. A systematic review. J Am Med Assoc 296:1094–1102.

Dunning D. 2006. Strangers to ourselves? Psychologist 19:600–603.

Eva KW, Regehr G. 2005. Self-assessment in the health professions: a reformulation and research agenda. Acad Med 80:S46–54.

Kruger J, Dunning D. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self assessments. J Per Soc Psychol 77:1121–1134.

Ward M, Gruppen L, Regehr G. 2002. Measuring Self-Assessment: current State of the Art. Adv Health Sci Educ 7:63–80.