Best Evidence Medical Education Systematic Review

The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice

Contents

The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice.

Short title: Self-assessment

Key Words: evidence based practice, professional practice, self-assessment, self-efficacy, self-evaluation programs, systematic review

Synopsis

Self-assessment is recognised as an integral component of a health professional's personal development. However, self-assessment skills are rarely taught and the ability to self-assess is seldom tested. This systematic review updates and considers the evidence for the validity of self-assessment since Gordon's comprehensive review in 1991.

Dates of the Review

Review commenced March 2004.
Literature search January 1990 to February 2005 (with update February 2006).
Analysis completed January 2007.

Topic Review Group Membership

Lead reviewer and contact for correspondence: Dr Brian McKinstry, University of Edinburgh.

Order of authorship:
Mr Iain Colthart, NHS Education for Scotland
Dr Gellisse Bagnall, NHS Education for Scotland
Dr Alison Evans, University of Leeds
Ms Helen Allbutt, NHS Education for Scotland
Mr Alex Haig, NHS Education for Scotland
Dr Jan Illing, University of Newcastle
Dr Brian McKinstry, University of Edinburgh

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.

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.

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

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

Executive summary

Health professionals are increasingly expected to identify their own learning needs as part of regulatory requirements of good professional practice. The ability to evaluate one's own strengths and weaknesses in terms of clinical knowledge and clinical skills, however, cannot be assumed. Previous research in health care education has cast doubt over the accuracy of self-assessing one's own performance compared with an external measure but some studies have also been marred by poor methodological quality.

In order to determine whether specific methods of self-assessment can lead to changes in learning activity or clinical practice, we undertook a systematic review of the health professions' literature. The method adopted followed a well described framework and included a comprehensive search of all sources pertinent to health professional education. A large number of papers were identified but only a small proportion of these were deemed to be relevant or of sufficient rigour to be included in a narrative review of results.

The majority of papers we reviewed addressed the accuracy of self-assessment in a clinical training context against some external standard. Very few studies 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. The review was therefore largely unable to answer the specific research questions and provide a solid evidence base for effective self-assessment. There was, however, some evidence that practical skills may be better self-assessed than knowledge and that accuracy of self-assessment may be enhanced by increasing the learner's awareness of the standard to be achieved. The review included several studies that found over-estimation of competence by poor performers. This finding has implications for practice and is worthy of further study. The acceptability of self-assessment as an appropriate educational activity was seldom explored in the literature and the dearth of robust qualitative research is of particular concern in this field.

If self-assessment is to remain the cornerstone of continuing professional development in the health care professions, we need to have a greater understanding of what forms of self-assessment may be useful in determining learning needs and what impact these have on future learning activities. In setting appropriate goals for learning, professionals need to be aware of the limitations of self-assessment and the need to use information from a range of sources to provide broader, more holistic assessments of competence in health care practice.

Abstract

Title: The effectiveness of self-assessment on the identification of learner needs, learner activity and impact on clinical practice: final report on BEME systematic review.

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.

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 (United Kingdom Central Council for Nursing, Midwifery and Health Visiting,1999; American Medical Association, 1992; 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 and 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 recognised 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 supervised 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 paper (1999). They set out to test in a non-clinical context the hypotheses that "incompetent individuals have more difficulty recognising 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. in press) 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:

- 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?

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 1. Although review papers were not used in answering our research questions, we have referred to relevant reviews in our discussion.

Box 1

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

1. Types of studies – research designs

All research designs were considered (see Box 2). 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 – see coding sheet) 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 2

Research designs

Type of study

Pilot	Prospective
Qualitative	Retrospective
Quantitative	Randomised 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-randomised 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 3. We also included outcome measures of accuracy of self-assessment and the factors influencing self-assessment. Additional predetermined and unintended outcomes were also accepted.

Box 3

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.

Search Strategy

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

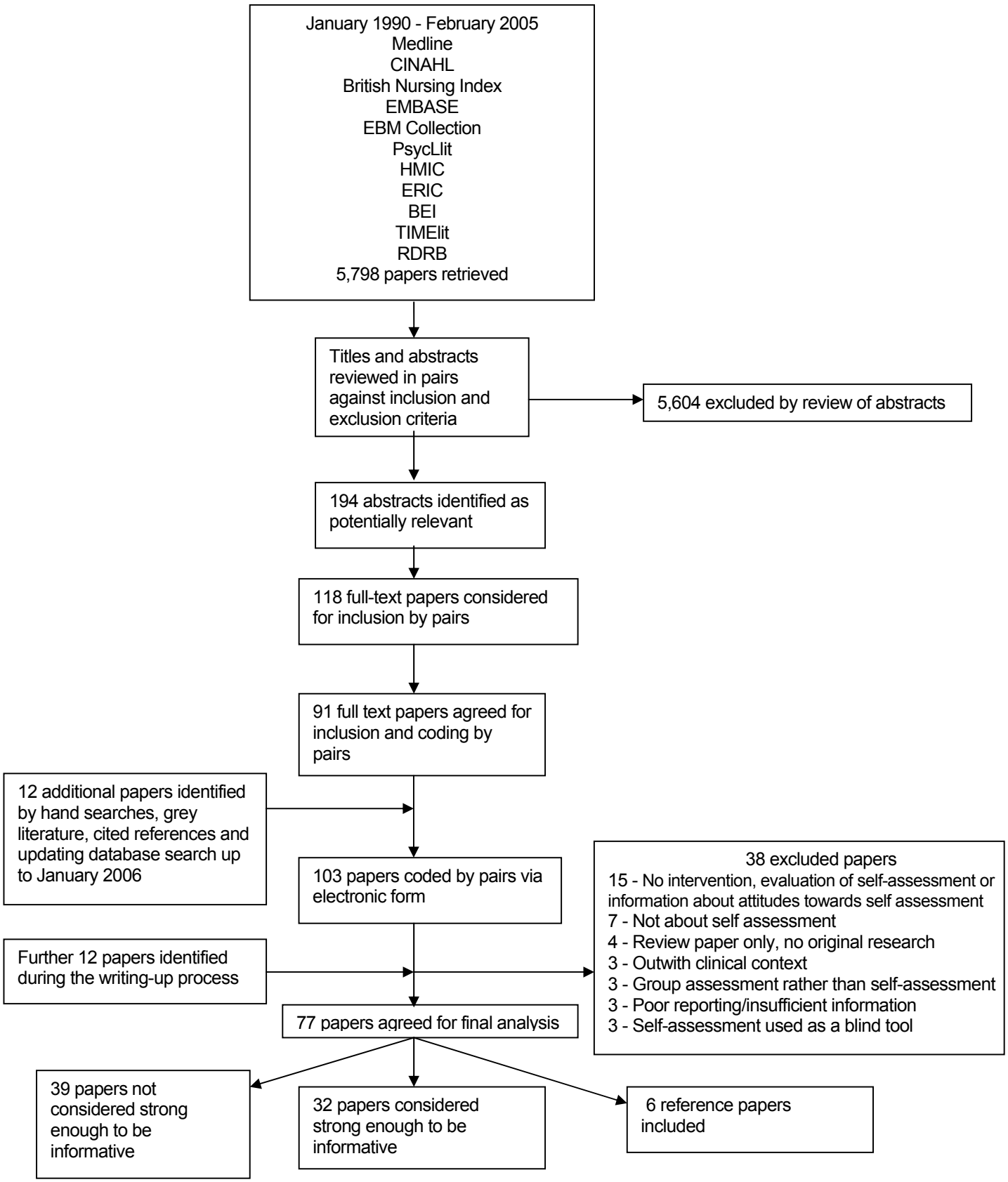
The database search covered all relevant health as well as educational databases, and included: Medline, CINAHL, BNI, Embase, EBM Collection, Psychlit, HMIC, ERIC, BEI, TIMElit and RDRB. The strategies were designed and tested for maximum sensitivity to ensure no potentially relevant papers were missed. 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 Appendix 1.

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 and 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 recognised 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 summarised in Figure 1 below. 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).

Figure 1: Flowchart of Search and Selection Strategy



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 (table 2).

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
- 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 (below).

Box 4a

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.

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. Again papers with grades 1 and 2 were discarded.

Box 4b

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.

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.

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 summarised 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 recognised 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 summarised 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 5 - Distribution of papers

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

Included papers are summarised in Table 1, and the excluded papers are listed in Table 2 with reasons for exclusion.

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 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 generalisability 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 and 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 video-taped 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 summarises 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 and Dunning (1999). There is a separate section later in this section specifically addressing insight. With reference to Kruger and 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 (=/- 0.8) and 3.5 (=/- 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 and 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 and Dunning, 1999) explored the hypothesis that incompetent students over-estimate their ability because their incompetence denies them the ability to recognise 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 standardised patient (SP) were compared with those of the standardised 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

recognise 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 and 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 (Reiter et al., 2002, Sullivan et al.,1999)
- 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 and 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 ($p < .05$) 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), summarised 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 ($p<or = 0.001$). 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%, $p < .001$). 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 personalised 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 finding in 1991 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 over-estimate 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; Bandura 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" (p548). 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 (van der Vleuten & Swanson,1990, Carline et al,.1989; 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, randomised 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.

Bibliography of Reviewed Papers

Citations that were strong enough to be informative

ANTONELLI, M.A. (1997) Accuracy of second-year medical students' self-assessment of clinical skills. *Academic Medicine*, 72(10 Suppl 1): pp. S63-65.

BRYAN, R.E., KRYCH, A.J., CARMICHAEL, S.W., VIGGIANO, T.R., PAWLINA, W. (2005) Assessing professionalism in early medical education: Experience with peer evaluation and self-evaluation in the gross anatomy course. *Annals Academy of Medicine Singapore*, 34, pp. 486-491.

EDWARDS, R., KELLNER, K., SISTROM, C., MAGYARI, E. (2003) Medical student self-assessment of performance on an OG clerkship. *American Journal of Obstetrics and Gynecology*, 188 (4), pp. 1078-1082.

ERICSON, C., CHRISTERSSON, C., MANOGUE, M., ROHLIN, M. (1997) Clinical Guidelines and Self Assessment in Dental Education. *European Journal of Dental Education*, 1, pp. 123-128.

EVA, K.W., CUNNINGTON, J.P.W., REITER, H.I., KEANE, D.R., NORMAN, G.R. (2004) How Can I Know What I Don't Know? Poor Self Assessment in a Well-Defined Domain. *Advances in Health Sciences Education*, 9, pp. 211-224.

EVANS, A.W., LEESON, R.M., NEWTON JOHN, T.R., PETRIE, A. (2005) The influence of self-deception and impression management upon self-assessment in oral surgery. *British Dental Journal*, 198(12), pp. 765–769.

FITZGERALD, J., WHITE, C., GRUPPEN, L. (2003) A longitudinal study of self-assessment accuracy, *Medical Education*, 37, pp. 645-649.

FITZGERALD, J.T., GRUPPEN, L.D., WHITE, C.B. (2000) The Influence of Task Formats on the Accuracy of Medical Students' Self-assessments. *Academic Medicine*, 75 (7), pp. 737-741.

GRUPPEN, L., WHITE, C., FITZGERALD, T., GRUM, C., WOOLLISCROFT, J. (2000) Medical Students Self-assessments and their Allocations of Learning Time. *Academic Medicine*. 75(4), pp. 374-379.

HERBERT, W.N., MCGAGHIE, W.C., DROEGEMEULLER, W., RIDDLE, M.H., MAXWELL, K.L. (1990) Student evaluation in obstetrics and gynecology: self vs departmental assessment. Obstetrics and Gynecology, 76(3 pt1), pp. 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. *Academic Medicine*, 76(10), pp. S87-89.

JOHNSON, D., CUJEC, B. (1998) Comparison of self, nurse and physician assessment of residents rotating through an ICU. *Critical Care Medicine*, 76(11), pp. 1811-1816.

LANE, J.L., GOTTLIEB, R.P. (2004) Improving the interviewing and self-assessment skills of medical students: is it time to readopt videotaping as an educational tool? *Ambulatory Paediatrics*, 4 (3), pp. 244-248.

LEOPOLD, S.S., MORGAN, H.D., KADEL, N.J., GARDNER, G.C., SCHAAD, D.C., WOLF, F.M. (2005) Impact of educational intervention on confidence and competence in the performance of a simple surgical task. *Journal of Bone and Joint Surgery*, 87A(5), pp. 1031-1037.

MANDEL, L.S., GOFF, B.A., LENTZ, G.M. (2005) Self-assessment of resident surgical skills: Is it feasible? *American Journal of Obstetrics & Gynecology*, 193(5), pp. 1817-1822.

MARTIN, D., REGEHR, G., HODGES, B., McNAUGHTON, N. (1998) Using videotaped benchmarks to improve the self-assessment ability of family practice residents. *Academic Medicine*, 73, pp. 1201–1206.

MILLIS, S.R., JAIN, S.S., EYLES, M., TULSKY, D., NADLER, S.F., FOYE, P.M., ELOVIC, E., DELISA, J.A. (2002) Assessing Physicians' Interpersonal Skills: Do patients and physicians see eye to eye? *American Journal of Physical Medicine & Rehabilitation*, 81(12), pp. 946–951.
MINTER, R.M., GRUPPEN, L.D., NAPOLITANO, K.S., GAUGER, P.G. (2005) Gender differences in the self-assessment of surgical residents. *American Journal of Surgery*, 189(6), pp. 647-650.
PARADISE, J.E., FINKEL, M.A., BEISER, A.S., BERENSON, A.B., GREENBERG, D.B., WINTER, M.R. (1997) Assessments of girls' genital findings and the likellihood of sexual abuse: agreement among physicians self-rated as skilled. *Archives of Pediatrics and Adolescent Medicine*, 151(9), pp. 883-891.
PARKER R.W., ALFORD C., PASSMORE C. (2004) Can Family Medicine Residents Predict Their Performance on the In-Training Examination? *Family Medicine*, 36(10), pp. 705-709.
REITER, H., EVA, K., HATALA, R., NORMAN, G. (2002) Self and Peer Assessment in Tutorials: Application of a Relative-ranking Model, *Academic Medicine*, 77(11), pp. 1134-1139.
RUDY, D.W., FEJFAR, M.C., GRIFFITH, C.H., WILSON, J.F. (2001) Self and peer assessment in a first year communication and interviewing course. *Evaluation and the Health Professions*, 24(4), pp. 436-445.
SOMMERS, P.S., MULLER, J.H., OZER, E.M., CHU, P.W. (2001) Perceived self-efficacy for performing key physician faculty functions - a baseline assessment of participants in a one-year faculty development program. *Academic Medicine*, 76(10), pp. S71-73.
SULLIVAN, M.E., HITCHCOCK, M.A., DUNNINGTON, G.L. (1999) Peer and Self Assessment During Problem Based Tutorials. *American Journal of Surgery*, 177, pp. 266-269.
TRACEY, J., ARROLL, B., BARHAM, P., RICHMOND, D. (1997) The validity of general practitioners self assessment of knowledge: cross sectional study, *British Medical Journal*, 315, pp. 1426-1428.
WARD, M., MACRAE, H., SCHLACHTA, C., MAMAZZA, J., POULIN, E., REZNICK, R., REGEHR, G. (2003) Resident self-assessment of operative performance. *American Journal of Surgery*, 185, pp. 521-524.
WEISS, P.M., KOLLER, C.A., HESS, L.W., WASSER, T. (2005) How do medical student self-assessments compare with their final clerkship grades? *Medical Teacher*, 27(5), pp. 445-449.
WILKERSON, L., LEE, M., HODGSON, C.S. (2002) Evaluating curricular effects on medical students knowledge and self-perceived skills in cancer prevention. *Academic Medicine*, 77(10), pp. S51–53.
WOODS, R., McCARTHY, T., BARRY, M.A., MAHON, B. (2004) Diagnosing Smallpox: Would you know it if you saw it? *Biosecurity and Bioterrorism: Biodefense Strategy, Practice and Science*, 2(3), pp. 157-163.
WOOLLISCROFT, J.O., TENHAKEN, J., SMITH, J., CALHOUN, J.G. (1993) Medical students' clinical self-assessments: comparisons with external measures of performance and the students self-assessments of overall performance and effort. *Academic Medicine*, 68(4), pp. 285-294.
YOUNG, J.M., GLASZIOU, P., WARD, J.E. (2002) General practitioners' self-ratings of skill in evidence based medicine: validation study. *British Medical Journal*, 324, pp. 950-951.
ZONIA S.L., STOMMEL, M. (2000) Interns Self-evaluations Compared with Their Faculty's Evaluations. *Academic Medicine*, 75(7). pp 742.

Citations that were not strong enough to be informative

ADRIAANSEN, M.J.M., VAN ACHTERBERG, T. (2004) A test instrument for palliative care. *International Journal of Nursing Studies*, 41, pp. 107-117.

ALNASIR, F., GRANT, N. (1999) Student self-assessment in a community-based clinical clerkship in family medicine: a preliminary report. *Education for Health*, 12(2), pp.161-166.

BAKKEN, L.L., SHERIDAN, J., CARNES M. (2003) Gender differences among physician-scientists in self-assessed abilities to perform clinical research. *Academic Medicine*, 78 (12), pp. 1281-1286.

BARNSLEY, L., LYON, P.M., RALSTON, S.J., HIBBERT, E.J., CUNNINGHAM, I., GORDON, F.C., FIELD, M.J. (2004) Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. *Medical Education*, 38, pp. 358–367.

BIERNAT, K., SIMPSON, D., DUTHIE, E., BRAGG, D., LONDON, R. (2003) Primary Care Residents Self-assessment Skills in Dementia. *Advances in Health Sciences Education*, 8, pp. 105–110.

CHUR-HANSEN, A. (2000). Medical students' essay-writing skills: criteria-based self-and tutor-evaluation and the role of language background. *Medical Education*, 34, pp. 194–198.

CLARIDGE, J.A., CALLAND, J.F., CHANDRASEKHARA, V. (2003) Comparing resident measurements to attending surgeon self-perceptions. *American Journal of Surgery*, 185, pp. 323-327.

COUTTS, L., ROGERS, J. (1999) Student Assessment and Standardized Patients - Will the Questions Never End? *Academic Medicine*. 74(10), pp. S128-130.

DAS, M., MPOFU, D., DUNN, E., LANPHEAR, J.H. (1998) Self and tutor evaluations in problem-based learning tutorials: is there a relationship? *Medical Education,* 32, pp. 411-418.

DAVIS, J.D. (2002) Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. *The American College of Obstetrics and Gynecology*, 99(4), pp. 647-651.

ELLIOTT, N., HIGGINS, A. (2005) Self and peer assessment - does it make a difference to student group work. *Nurse Education in Practice,* 5(1), pp. 40-48.

EVANS, A., MCKENNA, C., OLIVER, M. (2001) Towards better understanding of self-assessment in oral and maxillofacial surgery. *Medical Education*, 35(11), pp. 1077.

FARNHILL, D., HAYS, S.C., TODISCO, J. (1997) Interviewing skills: self-evaluation by medical students. *Medical Education,* 31(2), pp. 122-127.

FINCHER R., LEWIS, L. (1994) Learning, experience, and self-assessment of competence of third-year medical students in performing bedside procedures. *Academic Medicine*, 69(4), pp. 291-295.

FINCHER, R.M., 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. *Academic Medicine*, 68(2), pp. S47-S50.

FOX, R.A., INGHAM CLARK, C.L., SCOTLAND, A.D., DACRE, J.E. (2000) A study of pre-registration house officers' clinical skills. *Medical Education*, 34, pp. 1007-1012.

FRYE, A.W., RICHARDS, B.F., BRADLEY, E.W., PHILP, J.R. (1991) The consistency of students self-assessments in short essay subject matter examinations. *Medical Education*, 25, pp. 310-316.

HARRINGTON, J.P., MURNAGHAN, J., REGEHR, G. (1997) Applying a relative ranking model to the self -assessment of extended performances. *Advances in Health Sciences Education*, 2, pp. 17-25.

HARTMAN, S.L., NELSON, M.S. (1992) What we say and What we do: Self-reported teaching behavior versus performances in written simulations among medical school faculty. *Academic Medicine*, 67(8), pp. 523-527.

HOPPE, R.B, FARQUHAR, L.J., HENRY, R., STOFFELMAYR, B. (1990) Residents' Attitudes towards and Skills in Counseling: Using Undetected Standardised Patients. *Journal of General Internal Medicine*, 5, pp 415-420.
KAISER, S., BAUER, J. (1995) Checklist Self-Evaluation in a Standardized Patient Exercise. *The American Journal of Surgery*, 169 (April), pp. 418-420.
LEVINSON, W., GORDON, G., SKEFF, K. (1990) Retrospective versus actual pre-course self-assessments. *Evaluation and the Health Professions*, 13(4), pp. 445–452.
LIND, D.S., REKKAS, S., BUI, V., LAM, T., BEIRLE, E., COPELAND, E. (2002) Competency-Based Student Self-Assessment on a Surgery Rotation. *Journal of Surgical Research*, 105, pp. 31-34.
LORENZ, R., GREGORY, R.P., DAVIS, D.L. (2000) Utility of a brief self-efficacy scale in clinical training program evaluation. *Evaluation and the Health Professions*, 23(2), pp. 182- 193.
MacDONALD, J., WILLIAMS, R.G., ROGERS, D.A. (2003) Self-assessment in simulation-based surgical skills training. *The American Journal of Surgery*, 185, pp. 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. *European Journal of Dental Education*, 8, pp. 97-104.
MILLER, P.J. (1999) The agreement of peer assessment and self-assessment of learning processes in problem-based learning. *Journal of Physical Therapy Education*, 13(2), pp. 26-30.
MURDOCH-EATON, D. (2002) Reflective Practice Skills in Undergraduates. *Academic Medicine*, 77(7), pp. 734.
MURDOCK, J.E., NEAFSEY, P.J. (1995) Self-efficacy measurements: an approach for predicting practice outcomes in continuing education? *Journal of Continuing Education in Nursing*. 26(4), pp. 158-165.
PIERRE, R.B., WIERENGA, A., BARTON, M., THAME, K., BRANDAY, J.M., CHRISTIE, C.D. (2005) Student self-assessment in a paediatric objective structured clinical examination. *West Indian Medical Journal*, 54(2), pp. 144-148.
REES, C., SHEPHERD, M. (2005) Students' and assessors' attitudes towards students' self-assessment of their personal and professional behaviours. *Medical Education*, 39, pp. 30-39.
SKLAR, D.P., TANDBERG, D. (1995) The value of self-estimated scholastic standing in residency selection. *The Journal of Emergency Medicine*. 13(5), pp. 683-685.
STEWART, J., OHALLORAN, C., BARTON, J.R., SINGLETON, S.J., HARRIGAN, P., SPENCER, J. (2000) Clarifying the concepts of confidence and competence to produce appropriate self-evaluation measurement scales. *Medical Education*, 34, pp. 903-909.
TOUSIGNANT, M., DESMARCHAIS, J.E. (2002) Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study. *Advances in Health Sciences Education*, 7, pp. 19-27.
WAGNER, C., ABHOLZ, H.H. (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]. *Zeitschrift fur Allgemeinmedizin*, 80, pp. 150-152.
WEERAKOON, P.K., FERNANDO, D.N. (1991) Self-evaluation of skills as a method of assessing learning needs for continuing education. *Medical Teacher*, 13(1), pp. 103.
WETHERELL, J., MULLINS, G., HIRSCH, R. (1999) Self-assessment in a problem-based learning curriculum in dentistry. *European Journal of Dental Education*, 3, pp. 97-105.
WINDISH, D.M., KNIGHT, A.M., WRIGHT, S.M. (2004) Clinician Teachers Self-assessment versus learners' perceptions. *Journal of General Internal Medicine*, 19, pp. 554-557.
ZIJLSTRA-SHAW, S., KROPMANS, T.J., TAMS J. (2005) Assessment of professional behaviour - a comparison of self-assessment by first year dental students and assessment by staff. *British Dental Journal*, 198(2), pp. 165-171.

Additional References

AMERICAN MEDICAL ASSOCIATION (1992) 'Mirror, mirror: medicine enters the era of self-assessment'. *ALP Observer*, 12, pp. 17.

AMERY, J. , LAPWOOD, S. (2004) A study into the educational needs of children's hospice doctors: a descriptive quantitative and qualitative survey. *Palliative Medicine*, 18, pp. 727-733.

BANDURA, A. (1977) Self-efficacy: Toward a unifying theory of behavioural change. *Psychological Review*, 84, pp. 191-215.

BANDURA, A. (1982) Self-efficacy mechanism in human agency. *American Psychologist,* 37, pp. 122-147.

BANDURA, A. (1986) Social Foundations of Thought and Action: a social cognitive theory. (Englewood Cliffs, NJ, Prentice-Hall)

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).

BOUD, D. (1995) Enhancing Learning through Self Assessment (London, Kogan Page)

BRITISH MEDICAL ASSOCIATION (2003) Appraisal, a guide for medical practitioners. (London, BMA Publications)

CARLINE, J.D.,WEINRICH, M.D., RAMSEY, P.G.(1989) Characteristics of ratings of physician competence by professional associates. *Evaluation and the Health Professionals,* 12, pp. 409-23.

DORNAN, T, LITTLEWOOD, S., MARGOLIS, S.A., SCHERPBIER, A., SPENDER, J., YPINAZAR, V. (2006). How can experience in clinical and community settings contribute to early medical education? A BEME systematic review. *Medical Teacher*, 28 (1), pp. 3-18.

FRENCH MEDICINE ASSOCIATION cited in: Bruneau, C., Lachenaye-Llanas, C. (2002) The French Accreditation System. *Clinical Governance Bulletin*, 3, pp. 8-10.

GENERAL MEDICAL COUNCIL (2002) 'Tomorrow's Doctors: recommendation on undergraduate medical education', 2nd Edition, (London, GMC).

GORDON, M.J. (1991) A Review of the Validity and Accuracy of Self-assessments in Health Professions Training. *Academic Medicine*, 66, pp. 762-769.

GORDON, M.J. (1992) Self-assessment Programs and Their Implications for Health Professions Training. *Academic Medicine*, 67, pp. 672-679.

HAIG, A., DOZIER, M. (2003) BEME Guide No.3: systematic searching for. evidence in medical education - Part 1: sources of. information. *Medical Teacher,* 25(4), pp. 352-363

HAMMICK, M., FREETH, D., KOPPEL, I., REEVES, S., BARR, H. A best evidence systematic review of interprofessional education. *Medical Teacher*. In press.

JAMTVEDT, G., YOUNG, J.M., KRISTOFFERSEN, D.T., O'BRIEN, M.A., OXMAN, A.D. (2006). Audit and feedback: effects on professional practice and health care outcomes. *Cochrane Database of Systematic Reviews*, Issue 2. Art. No.: CD000259. DOI: 10.1002/14651858.CD000259.pub2

KIRKPATRICK, D.L. (1967) Evaluation of training, in: Craig, R. and Bittel, L. (Eds) *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.userpage.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, C.P.M. , SWANSON, D.B. (1990) Assessment of clinical skills with standardized patients: state of the art. *Teaching and Learning in Medicine*, 2, pp. 58-76.
WILLIAMS, R.G., KLAMEN, D.A., McGAGHIE, W.C. (2003) Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine.* 15(4), pp. 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, J.A., VERHULST, S.J., BARROWS, H.S. (2005) Self-assessment in medical practice: a further concern about the conventional research paradigm. *Teaching and Learning in Medicine.* 17(3), pp. 200–201.
DAVIS, D.A., MAZMANIAN, P.E., FORDIS, M., VAN HARRISON, R., THORPE, K.E., PERRIER, L. (2006) Accuracy of physician self-assessment compared with observed measure of competence. A systematic review. *Journal of the American Medical Association,* 296 (9), pp. 1094-1102.
DUNNING, D. (2006) Strangers to ourselves? *Psychologist*, 19(10), pp. 600-603.
EVA, K.W., REGEHR, G. (2005). Self-assessment in the health professions: a reformulation and research agenda*. Academic Medicine*, 80(10) (Suppl), pp. S46–54.
KRUGER, J., DUNNING, D. (1999) Unskilled and unaware of it: how difficulties in recognizing ones own incompetence lead to inflated self assessments. *Journal of Personality and Social Psychology*, 77 (6), pp. 1121-1134.
WARD, M., GRUPPEN, L., REGEHR, G. (2002) Measuring Self-Assessment: Current State of the Art. *Advances in Health Sciences Education*, 7, pp. 63-80.

BEME Disclaimer

BEME review results are, necessarily, interpreted in light of individual perspectives and circumstances. The conclusions presented in this review are the opinions of review authors. Their work has been supported by BEME but their views are not necessarily shared by all BEME members.

The aim of BEME is to make the results of research into the effectiveness of educational interventions available to those who want to make more informed decisions. This information is an essential contribution to the process of deciding whether to adopt a particular educational intervention or not. Information and the assessment of needs, resources and values; as well as judgements about the quality and applicability of evidence are equally important. It is unwise to only rely on evidence about the impact of a particular educational intervention. Understanding learning process for the students in your context, knowledge of past success and failures and how educational interventions work are all vital. BEME does not accept responsibility for the results of decisions made on the basis of a BEME Review.

Appendices

Appendix 1: Search Strategy

Search Strategy

A standard search strategy was designed to identify self-assessment interventions and then adapted for each electronic database. The following is the Medline strategy:

1 SELF ASSESSMENT/
2 SELF EFFICACY/
3 SELF_EVALUATION PROGRAMS/
4 (self adj (assess$ or evaluat$ or rat$ or grad$ or apprais$)).mp.
[mp=title, original title, abstract, name of substance, mesh subject heading]
5 EDUCATION, MEDICAL, GRADUATE/ or EDUCATION, PHARMACY, CONTINUING/or EDUCATION, DISTANCE/ or EDUCATION, NURSING, DIPLOMA PROGRAMS/ or EDUCATION, PREMEDICAL/ or EDUCATION, NURSING, GRADUATE/ or EDUCATION, PROFESSIONAL/ or EDUCATION, NURSING, CONTINUING/ or EDUCATION, NURSING/ or EDUCATION, VETERINARY/ or EDUCATION, PUBLIC HEALTH PROFESSIONAL/ or EDUCATION, DENTAL, GRADUATE/ or EDUCATION, PHARMACY, GRADUATE/ or EDUCATION, CONTINUING/ or EDUCATION, MEDICAL/ or EDUCATION, MEDICAL, UNDERGRADUATE/ or EDUCATION, PHARMACY/ or EDUCATION, NURSING, ASSOCIATE/ or EDUCATION, NURSING, BACCALAUREATE/ or EDUCATION, GRADUATE/ or EDUCATION, DENTAL/ or EDUCATION, PROFESSIONAL, RETRAINING/ or EDUCATION, MEDICAL, CONTINUING/ or EDUCATION, DENTAL, CONTINUING/
6 exp PROFESSIONAL COMPETENCE
7 ed.fs.
8 exp LEARNING/
9 6 and (7 or 8 or 5)
10 5 or 7 or 8 or 9
11 1 or 2 or 3 or 4
12 10 and 11

Appendix 2: Coding Sheet

The electronic coding sheet can be accessed at:

http://134.36.210.98/cgi-bin/survey/survey/24

Appendix 3: Contact for Topic Review Group Members

Dr Brian McKinstry, Senior Researcher, Department of General Practice, University of Edinburgh, 20 West Richmond Street, Edinburgh, EH8 9DX.

brian.mckinstry@ed.ac.uk

Table 1 Summary of Papers Included for Analysis

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
1.	Antonelli MA (1997)	To determine if giving students specific skill objectives would lead to accurate self-knowledge concerning performance of the specific skills	USA Medical students (second year) N = 87	Comparative study Medical students predicted performance on 78 item checklist using 5 point scale. And overall performance on 4 point scale form unsatisfactory to honours. This was compared with preceptor marking, written and final exam grade. Perception of self-assessment skill was measured on a 4 point scale.	1. Do self-assessment scores correlate with preceptor ratings and examination grade? 2. Can students identify if they are accurate self-assessors?	63 students took part but only 47 had complete data. Self assessment of course performance was correlated with the exam grade ($r > 0.4$, $p = 0.008$: $n=50$) and preceptor grade ($r=.031$, $p = 0.03$). However there was poor correlation with specific items on examination and history taking. Higher achieving students graded their overall performance higher than lower achieving student, but they already knew 2/3s of their mark. There were no gender differences in ability to self-assess. There was no correlation between perception of self assessment skill and accuracy of self assessment.	Students were better at making a global assessment of their performance than on specific skills (however they already had a large degree of feedback on their overall grade knowing 2/3 of the mark). Students were not able to identify if they could they were accurate self assessors.
2.	Bryan RE, Krych AJ, Carmichael SW, Viggiano TR, Pawlina W (2005)	To determine if peer evaluation and self-evaluation used in conjunction and implemented early in the medical curriculum can serve as useful tools to assess and provide feedback re professional behaviour in first year medical students.	USA Medical students n = 213	Comparative Study Over a period of five years medical students evaluated themselves and their peers during the Gross and Developmental Anatomy Course. Numerical evaluations and written comments were statistically analysed within established categories of professionalism and correlated with academic performance, gender, and peer-rating and self-rating.	What is the correlation between academic performance, gender, peer and self ratings on the first year gross and developmental anatomy course?	Peer evaluation versus self evaluation: 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, $p<0.001$. Academic performance: Linear regression indicated a slight positive correlation between the final grade and the total number of positive comments received ($r=0.22$, $p<0.001$). Gender differences: Males received significantly more positive comments than females on peer evaluations (mean, +/- sd) of 9.1 (+/- 2.5) and 8.4 (+/- 2.0) respectively ($p=0.025$) and were rated higher than	Results confirm those of Rudy et al. that peer provided positive comments correlated with higher peer evaluation scores but not with self-evaluation scores. Also reproduced previous findings that students with higher grades underestimated their own performance while those doing poorly tended to over-estimate it. These data suggest that similar to previously drawn conclusions peer assessment skills may not transfer to self assessment skills. Despite students willingness to evaluate their peers in an objective fashion, their self–evaluation is much more critical.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
						females on peer provided numerical rating (mean, +/- sd) of 4.4 (+/- 0.5) and 4.2 (+/- 0.5) respectively (p=0.02). Males also rated themselves more highly than did females (mean, +/- sd) 3.7 (+/- 0.8) and 3.5 (+/- 0.9) respectively (p=0.04).	
3.	Edwards R, Kellner K, Sistrom C, Magyari E (2003)	To evaluate the accuracy of medical student self-assessment of performance on an obstetrics and gynecology clerkship and to assess the influence of demographic and temporal factors on the accuracy of self-assessment.	USA Third year medical students participating in obstetrics & gynecology (OG) clerkship n = 1,152	Comparative study. Medical students predicted their exam and clerkship grades at beginning and end of OG clerkship. These were compared against actual grades.	1. What are the differences between predicted and actual final examination and clerkship grades as a measure of self-assessment accuracy? 2. Do demographic and temporal factors influence the accuracy of self-assessment?	Students were more likely to predict the grade they would receive for the clerkship than for the examination, both at the beginning (56% vs 31%, p <0.001) and the end (61% vs 32%, p<0.001) of the clerkship. Students were slightly more likely to predict correctly their clerkship grades at the end compared with the beginning of the clerkship (61% vs 56%, p=0.03). However they were no more likely to predict their examination grades at the end of their clerkship compared with the beginning (32% vs 31%, p=0.75). Students with higher grades tended to underestimate their performance, students with lower grades tended to overestimate their performance.	Overall third year medical students are moderately skilled at self-assessment. More than half of the students correctly predicted their clerkship grade, both at the beginning and end of the clerkship. However less than one third of them correctly predicted their examination grade regardless of the timing of the prediction. Men were 1.7 times more likely than women to overestimate their grades. Experience may improve the accuracy of self-assessment. Students may develop a more realistic appraisal of their abilities if they are given more specific and intensive feedback during their clerkship.
4.	Ericson C, Christersson C, Manogue M, Rohlin M (1997)	To compare students' self-assessment of their performance with those of their instructors using these guidelines.	Sweden Dental students n = 41 (35 participated)	Comparative study Students' performance in following clinical guidelines in treating 137 patients in cariology were scored on a 3 point scale (excellent, acceptable, unacceptable) by both students and instructors.	No clear research question.	Instructors and students agreed in 87% of their assessments. Students under- scored their performance more frequently (10%) and they over-scored it (3%) in comparison in to their instructors.	Under rating by students of their performance was seen to occur more frequently than over rating, which agrees with several other studies of students in the health professions.
5.	Eva KW, Cunnington JPW, Reiter HI,	To perform two studies which attempt to improve the fact that self-	Canada Medical students Study 1: n = 116 Study 2 : n = 265	Comparative studies Study 1: student's were asked to predict their ability in particular subject	Study 1. Will students' self-assessments be better correlated with	Study 1: the correlation between average self-assessment scores across all subject areas (i.e. factual and higher order) and PPI was near	Results of these studies present a strong challenge to the notion of a self-assessment skill. Despite favourable conditions (e.g. experience of self –assessment and

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
	Keane DR, Norman GR (2004)	ratings are generally poorly correlated with other performance measures.		areas before sitting a multiple choice exam, the Personal Progress Inventory (PPI). Predicted performance was compared against actual exam performance. Study 2: students were asked to predict how many questions they would answer correctly in the PPI; to assess their performance on the whole test; one half of the class was asked to do the assessment before the exam (prediction) and the other half after sitting the exam (postdiction).	actual performance for fact-based examination questions than for higher-order conceptual questions. Study 2. Q1 Will there be a large correlation between predictions and actual performance? Q2. Would senior students, who have had more self-assessment experience and a greater amount of previous feedback on the test, be more accurate in their predictions than junior students. Q3. Would students who have sat the exam (postdiction) be more accurate in self-assessment than those who had not yet sat it (prediction)?	zero for both types of questions. Study 2: On average students underestimated their performance but only slightly. Self-assessment improved after the experience of writing the test as all correlations showed positive changes between the prediction and postdiction groups. There was no evidence that performance in self-assessment improved with time in the programme. If anything correlations with percent correct declined with increasing seniority.	feedback) predictions of performance were modest at best. No evidence that as much as 2.5 years general experience in a self-assessment environment and specific feedback on 7 very similar tests provided any benefit to student's ability to predict their performance relative to their colleagues .i.e. no evidence that students improved their self-assessment skill after 2.5 years of schooling. This finding is consistent with other longitudinal studies of self-assessment abilities.
6.	Evans AW, Leeson RM, Newton John TR, Petrie A (2005)	To see if poor self assessment of surgical performance during removal of mandibular third molars is influenced by self-	United Kingdom Oral and maxillofacial surgeons n = 50	Comparative Study. The surgeons' surgical skills were assessed (by two assessors) and self-assessed using check-list and global rating scales. Post-operatively, surgeons completed	Is apparent over or under-marking of their own surgical skills by staff and postgraduates influenced by self-deception enhancement (sde)	Reliability between assessors was excellent for checklist (0.96) and global rating scales (0.89) and better than the reliability between assessors and surgeons (0.51 and 0.49). There was a statistically significant correlation ($r=0.45$, $p=0.001$ checklist, $r= 0.48$, $p<0.001$ global) between	The majority of surgeons scored themselves higher than their assessors did for surgical skill in removing a single mandibular third molar tooth. Impression management (the tendency to deliberately convey a favourable impression) may contribute to a surgeon's inaccurate self-reporting of performance. Lack of insight appears to be much less

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
		deception (lack of insight) and impression management (trying to convey a favourable impression).		validated deception questionnaires which measured both self-deception enhancement (lack of insight), and impression management (the tendency to deliberately convey a favourable impression). Reliability between assessors, and between assessors' and surgeons' self-assessments were calculated. Discrepancies between assessors' and surgeons' scores were correlated with surgeons' deception scores.	i.e. lack of insight or impression management (im) i.e. faking good?	over/under-rating of their surgical performance by surgeons and their impression management scores. No statistically significant correlation was found between this inaccuracy in self-assessment and surgeons' individual self-deception scores.	important as a contributing factor. The authors speculate that pressure to provide evidence of good performance may be encouraging surgeons to manage their image and over-score themselves.
7.	Fitzgerald JT, Gruppen LD, White CB (2000)	To examine the relationships between self assessment and task accuracy	USA Medical students n = 304	Comparative study. Students assessed their own performances on ten stations of a clinical examination. The examination used two formats: performance tasks (the examination or history taking of standardized patients) and cognitive tasks (interpreting vignettes or test results and then answering paper-and-pencil questions).	1. Does the accuracy of students' self-assessments vary across task formats? 2. Are students' self-assessments consistent within task formats?	Three measures of self-assessment accuracy were used: a bias index (average difference between the students' estimates of their performances and their actual scores), a deviation index (average absolute difference between estimate and actual score), and an actual score–estimate-of-performance correlation (the correlation between the estimate and actual scores). The student bias and deviation indices were similar on the cognitive and the performance tasks. The correlations also indicated similarity between the two types of tasks.	The results indicate that the format of the task does not influence students' abilities to self-assess their performances, and that students' self-assessment abilities are consistent over a range of skills and tasks.
8.	Fitzgerald J, White C, Gruppen L (2003)	To document the stability of self-assessment accuracy over time by comparing actual and	USA Medical students. N = 500 over three years (163 year 1; 169 year 2; 168 year 3)	Comparative study. Students assessed their performance on classroom examinations and objective structured clinical examination	Are self-assessment skills stable over time or are they learnable or modifiable?	The multivariate repeated measures analysis of variance indicated that all three self-assessment accuracy measures, self-estimates and performance scores changed over the course of the study. The bias	The means for the performance and self-assessment accuracy measures reflect a fairly high level of stability during the first three assessment periods. However, when self-assessment is required on a different type of task (the third year

Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
	estimated examination performance for three classes during their first three years of medical school		(OSCE) stations. Each self-assessment was then contrasted with their actual performance using idiographic (within-subject) methods to define three measures of self-assessment accuracy: bias (arithmetic differences of actual and estimated scores), deviation (absolute differences of actual and estimated scores), and covariation (correlation of actual and estimated scores). These measures were computed for four intervals over the course of 3 years. Multivariate analyses of variance and correlational analyses were used to evaluate the stability of these measures.		scores were negative (indicating an under-estimation of actual performance on average) for the first three periods but became positive in the third year. This suggests that on average the students overestimated their performance on the OSCE. Change patterns in the deviation and covariation values were similar to the bias measure. In the first three periods scores were relatively consistent but in the third year the deviation score increased from 7.8 to 12.9 while the mean covariation score decreased from 0.36 to 0.26. The same pattern of change was also evident in the actual self-estimates students provided and the actual performance, both of which showed a decrease in the third year. In terms of correlations between consecutive periods; the correlations between the three self-assessment accuracy measures indicated that the bias and deviation measures had a similar pattern. For both of these measures, the relative stability of students' self-assessment accuracy was moderately high from one period to the next in the first two years of medical school with correlations ranging from 0.46 to 0.69. However the correlations between the second year and third year periods were substantially lower.	OSCE), both student performance and self-assessment scores change. For the first time, students overestimated their performance. The increase in both the deviation and covariation scores suggests that self-assessment has become less accurate, which in turn suggests that the type of task or task experience might play a role in making self-assessment judgements. Results of this study indicate that medical student self-assessment accuracy is reasonably stable when compared with the stability of actual performance. There may be multiple explanations for the decline in self-assessment accuracy and actual performance between the classroom assessments of knowledge (in the first 2 years) and the clinical assessments of diagnostic and procedural skills (in the OSCE). One is task familiarity. Students who enter medical school have spent years taking paper and pencil examinations. When the task is one in which the students have had limited experience, self-assessment accuracy suffers, as does performance. An alternative explanation may be that self-assessing one's knowledge (as in the first year assessments) is a different process from self-assessing one's performance (as in the OSCE). It may be that self-assessment of knowledge requires dimensions and information that are different from those required in the self-assessment of performance.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
9.	Gruppen L, White C, Fitzgerald T, Grum C, Woolliscroft J (2000)	To investigate the impact of self-assessed diagnostic strengths and weaknesses on medical students' allocation of learning time.	USA Medical students. N = 107	Comparative study. Self-assessment at beginning and end of clerkship and estimates of time spent on topic at end of three months.	1. What is the relationship between students' initial self-assessment of diagnostic strengths and weaknesses in specific topics and allocations of educational time to same topics? 2. How do amounts of study time correlate with changes in self-assessed diagnostic capabilities over the course of the clerkship?	1. students' initial self-assessments of their individual diagnostic skills relevant to the 14 topics were not associated with variations in how they allocated their educational time (average tau-b coefficient = 0.06, SD = 0.24; 95% CI = 0.01 to 0.11). However there was , considerable individual variation in the strength of this relationship (tau-b range = -0.68 to 0.51); some students spent more time on topics in which they felt relatively weak (yielding a negative individual correlation), whereas others spent more time on topics in which they already felt strong (yielding a positive correlation). 2. The average correlation between each student's relative allocation of learning time among these complaints and the magnitude and direction of change in his or her self-assessed diagnostic skills was moderate and positive (mean coefficient = 0.25, SD = 0.20; 95% CI = 0.21 to 0.29). The positive correlation indicated that spending relatively more time learning about a given chief complaint resulted in a relatively greater increase in self-assessed diagnostic skill related to that complaint. Again, however, there was considerable individual variation in the strengths of this relationship (range = -0.23 to 0.89), indicating that it was considerably stronger for some students than for others.	Individual-level analyses indicated that, for the average student, self-assessed strengths and weaknesses did not correlate with allocation of educational time, but that time allocation was positively related to changes in self-assessed skill. Study suggests that students were in early stages of self directed learning and so are 'beginners' in this activity of self directed learning. Self assessment and self –directed learning are individual and idiosyncratic phenomena and may not lend themselves easily to generalisable relationships.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
10.	Herbert W, McGaghie W, Droegemeuller W, Riddle M, Maxwell K (1990)	1. to correlate students' self-evaluation with departmental evaluations used to judge performance 2. to assess correlation among grades assigned by different groups (faculty, residents) and different methods (clinical activities, written examinations, oral examinations) 3. to determine the effect of gender and previous clerkship experience on students' perceptions of their achievements or their actual clerkship grades.	USA Third year medical students in obstetrics and gynecology clerkship n = 142	Comparative study. Students completed forms for expected grade in several performance categories; also provided information on previous clerkship and gender. Student predicted expected grades and actual departmental grades were compared with attendings and residents with respect to overall grades. The correlation between self and departmental ratings across seven categories and the difference in average ratings between self, residents and attendings were analyzed..	How do students' self-assessments correlate with various other measures of their performance?	*S*tudents graded themselves lower than attending ratings and higher than actual resident ratings. There were positive correlations between self and departmental evaluations for clinical activities (0.32), written examination (0.33), and oral exams (0.30) (all significant $p<0.0001$). In comparison of both departmental ratings and self-ratings for all methods of evaluation, there were no differences attributable to student gender. In terms of previous clerkship experience there was no correlation between this and the grades given by the department for any of the seven evaluation parameters. This pattern was repeated for self-evaluation with the single exception of grades predicted by residents on the gynecology service.	Students' perceptions of their performance coincide with ratings by faculty and residents. Students anticipated lower grades from faculty than they actually received. There was a strong correlation between student and departmental evaluations in different methods of evaluation. Gender or more clerkship experience did not affect students' ability to self-assess.
11.	Hodges B, Regehr G, Martin D (2001)	To see if the work of Kruger and Dunning (those with the least skill may be most at risk of inaccurately assessing their abilities) could be demonstrated in family medicine residents.	USA First year family medicine residents n = 24	Comparative study. Comparison of pre and post video assessments for high and low performers. Intervention with standardised patient using self-assessment and expert assessment. Exposed to video with same standardised patient scenario, and	1. Were the lowest performers in the group the most inaccurate in self-assessment? 2. Would highest performers underestimate their skills, but recalibrate following exposure to performance of others?	Residents' raw scores were sorted by tertiles and those in the top and bottom tertiles initially scored themselves inaccurately compared to the expert's scores. Pre and post video assessments for both groups were then compared by a two tailed *t* test to see if residents in either group would change their scores significantly in the direction of the experts' scores. Using raw scores alone the changes were not	Residents in highest performance group were able to recalibrate their self-assessment more accurately when presented with benchmark videos. However such change in self-assessment was minimal or insignificant for lowest performance group of residents. Exposure to benchmark performances can lead to better self assessment, but not for residents in lowest performance group.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
				assessed it. Asked to rescore own self-assessment; all compared to expert ratings		significant (p=0.26 for low group and p=0.19 for high group). Using z scores re-scaled relative to the videos revealed another picture. The change in self-assessment toward the experts' scores was only marginally significant for the low performers (0.16 ± 0.28, p=0.074) but highly significant for the high performers (0.37 ± 0.24, p = 0.002). The difference in the amount of correction between the two groups (0.16 vs 0.37) was marginally significant (p=0.06).	
12.	Johnson D, Cujec B (1998)	To examine the hypothesis that resident ratings by self, attending physicians and nursing staff assess different aspects of performance and may not be closely correlated.	Canada Surgical, medical, anaesthesia and obstetrics residents. n = 60	Comparative study. Compare resident evaluations by self, nurses and attending physicians by means of exams, questionnaires (multiple choice exam), one of two subjective instruments (Global Rating Scale, Behaviourally Anchored Rating scale) and expert opinion	Do resident ratings by self, attending physicians and nursing staff assess different aspects of performance?	Physicians' evaluations correlated with the mcq test scores (Spearman's rho 0.3082, p=0.005, n=82), whereas neither self-evaluation (Spearman's rho 0.1124, p=0.65, n=42) nor nurses' evaluations (Spearman's rho 0.2060, p=0.069, n=79) had a significant correlation with test scores. Spearman's correlations were not significant for either overall competence or specific medical knowledge by any category of evaluator using the Global Rating Scale. Spearman's rho correlations and kappa statistic between the three types of evaluators (physicians, nurses and self) for each criterion of the Behaviorally Anchored Rating Scale demonstrated significant correlations between the ratings of physicians and nurses, except for the assessment of humanistic qualities.	Self-rating by residents did not correlate to mcq scores and differed in some criteria with physicians' or nurses' evaluations. Found many similarities and some differences between physicians' and nurses' evaluations of residents. Conclude that different categories of evaluators assess different aspects of performance. Assessment by a varied group of evaluators should be used when attempts to predict future practice are made.
13.	Lane JL, Gottlieb RP (2004)	To evaluate the effect of a videotaping programme on	USA Third year medical students. n = 60	Comparative study. A self-assessment manual (SAM), listing and explaining 21 core	How accurate are students in self–assessment and can these skills be	Students inaccurately self-assessed on the first video 14% of the time. The 6 poorest-performed core elements were the least accurately self-	The videotaping program improved students' interviewing and self-assessment skills and identified students with inflated views of their abilities.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
		medical students' interviewing and self-assessment skills		elements of the medical interview was developed. After reading the manual, students videotaped an interview and self-assessed their performances. Each student reviewed the videotape with a faculty member who also rated the performance. This process was repeated 1 week later. Changes in group performance, core element performance, and ability to self-assess after the intervention were evaluated	improved?	assessed. Lack of concordance between the global rating given by faculty and student identified all students with inflated self-assessment. One review session had a large effect on overall performance and interpersonal skills and a moderate effect on history-taking skills. A large effect on performance was seen for 3 core elements, a moderate effect for 12 elements, and a small effect for the remaining 6 elements. Performance of the core elements that needed improvement did improve in 74% of the students (P <.001). Students' overall ability to self-assess improved significantly (P <.01).	
14.	Leopold SS, Morgan HD, Kadel NJ, Gardner GC, Schaad DC, Wolf FM (2005)	To evaluate the interrelationship of level of confidence, background, education and skill in the performance of a simulated knee joint injection.	USA Medical doctors (n43), bone specialists (n3), advanced nurse practitioners (n35), physician's assistants (n12)	Comparative study. 93 practitioners attending a CPD course on outpatient management of musculoskeletal disorders 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	1. Is there a relationship between an individual's confidence in his/her ability to perform a task and his/her ability to perform the task competently? 2. Does psychomotor skills education improves the correlation between confidence and competence? 3. Are demographic variables associated with differences in the confidence-competence relationship? 4. Is there a	Before instruction, males were significantly more confident then females (6.32 vs 2.95 points p< 0.01) about performing the knee injection. There was no significant difference as assessed between men and women in their clinical skills performance. Physicians were also more confident than non-physicians (5.86 vs 2.78 points p < 0.01) although no significant difference was observed in performance between physicians and non-physicians. Confidence before instruction was inversely correlated with performance i.e. greater confidence was associated with poorer performance of the pre-instruction knee injection test. Confidence was directly correlated with self-assessed ability of their own performance. After instruction,	Men and physicians disproportionately overrated their skills both before and after training, a finding that worsened as confidence increased. Low intensity forms of instruction improve individuals' confidence, competence and s-a of their skill in performing a simple surgical task.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
				and after performances were graded by trained assessors expert in this field.	difference between different forms of instruction to improve competence?	physicians continued to self-report higher confidence scores than non-physicians (8.22 vs 6.98 points p =<0.01), a difference that was not associated with better performance. Females became more confident than males after instruction (8.77 vs 6.98 points p = < 0.01) and also had higher scores for performance. Hands-on teaching did not significantly improve performance more than a printed leaflet or video.	
15.	Mandel LS, Goff BA, Lentz GM (2005)	To examine obstetrics and gynecology residents' self-assessment of proficiency on a variety of surgical bench procedures and to compare their ratings with those ratings of trained faculty observers.	USA Surgical Residents n = 74 (92 approached)	Comparative Study Residents rated their overall open and laparoscopic skills on scale of 1-5 prior to the OSATS, then after each of 6 stations completed an overall performance scale for that procedure (1-5) and a 35 item global skills checklist (same for each station). Experienced faculty physicians also completed these ratings. Paired t tests were used to compare resident and faculty ratings. Comparisons by resident level or testing site were carried out using one-way analysis of variance, and additional analyses using Pearson bivariate correlations.	1.How do residents' ratings of their own surgical performance compare with those of trained faculty observers? 2. Are residents who score at the low end of the scale less accurate in their self-assessment?	Overall self-assessment and overall global scores were correlated significantly with the total overall and total global faculty ratings, and the residents' estimate of their open and laparoscopic skills before they started the OSATS (P < .001). Residents tended to rate themselves lower than did faculty members on both the individual tasks and composite ratings and global skills (P < .001). There was a high degree of correlation between resident and faculty ratings on specific tasks. Both self assessment and faculty scores increased with increasing experience level of the residents. Comparing overall faculty ratings versus self-assessment for the lowest-scoring residents (> or = 2 SDs below the mean) they found a significant correlation between the 2 assessments (P < .05) and very few instances in which the faculty rated the resident low but the resident rated him/herself high.	Residents can rate their overall open and laparoscopic skills, overall task-specific assessments and global skills with good reliability and validity. Although they tended to score themselves lower than did faculty observers, the correlations were high. They did not find that residents with poor skills were not aware of their deficiencies.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
16.	Martin D, Regehr G, Hodges B, McNaughton N (1998)	1.To explore certain methodologic issues that might have led to a systematic underestimate of the ability to self-assess. 2. To address methodologic and statistical problems of previous self-assessment studies by exposing participants to relevant standards, anchoring rating scales, and providing practice in the use of the assessment tool.	USA First and second year family practice residents n = 50	Comparative study Residents performed a patient interview. Following the interview the resident and two experts independently evaluated the resident's communication skills. The resident was then shown a video of 4 performances (ranging from poor to good) of the same scenario. The resident evaluated the communication skills displayed in each performance and then re-evaluated his or her own performance.	Can an individual's abilities to self assess be improved by providing video taped benchmarks for comparison?	The correlation between expert's evaluations and residents' self-evaluations was moderate immediately after the interview (r=0.38) but increased significantly after the residents viewed the videotape(r=0.52). This effect was more pronounced for first year residents (0.22 to 0.45) than for second year residents (0.53 to 0.65) although the difference was not significant. Post-hoc analysis revealed that neither initial nor post-benchmark self-assessment ability was related to the ability to accurately evaluate the benchmarks in a manner consistent with the experts.	Initial findings are consistent with past research in that residents correlations with the assessments of experts were poor to moderate. Correlations did improve with exposure to relevant benchmarks. Providing a set of benchmarks against which trainees can compare their own performances improves the ability to self-evaluate even if the benchmarks are not explicitly identified.
17.	Millis SR, Jain SS, Eyles M, Tulsky D, Nadler SF, Foye PM, Elovic E, DeLisa JA (2002)	To determine the level of agreement between standardized patient ratings and resident physician self-ratings of physician interpersonal skills and the level of agreement between faculty observer and standardized patient ratings of	USA Resident physicians in physical and rehabilitation medicine n = 25	Comparative study Structured clinical evaluation. Residents conducted a 10 minute interview of a standardized patient to obtain a history. A resident physician, a standardized patient, and a faculty observer rated the resident physician's interpersonal skills immediately after the interview. The main outcome measure was a	1.What is the level of agreement between standardized patient ratings and resident physician self-ratings of physician interpersonal skills 2. What is the level of agreement between faculty observer and standardized patient ratings of resident physicians'	There was a low level of agreement between standardized patient ratings and the resident physicians' self-ratings of interpersonal skills (Lin's concordance coefficient, r_c = 0.11, P = 0.58). Conversely, there was a statistically significant degree of agreement between the standardized patient and faculty observer ratings of resident physician interpersonal skills (r_c = 0.50, P = 0.006).	Some resident physicians have significant difficulty accurately assessing how well they communicate with patients. Physicians in training rarely get feedback regarding their interpersonal skills and may have difficulty using social comparison. Conversely, standardized patients and faculty observers may have insight into interpersonal skills about which resident physicians are unaware.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
		resident physicians' interpersonal skills.		modification of the patient assessment measure from the American Board of Internal Medicine, a 9-item rating scale assessing communication (score range, 9–45).	interpersonal skills?'		
18.	Minter RM, Gruppen LD, Napolitano KS, Gauger PG (2005)	To determine if the gender differences in self-perception identified in medical students also exist in male and female surgical residents	USA Clinical general and plastic surgery residents n = 39	Comparative Study The difference in ordinal values from faculty and resident self-evaluations was calculated for each general and plastic surgery resident. Objective external performance measures (the mean faculty value for each competency and the resident's score on the American Board of Surgery In–Training Examination) were compared for female and male residents.	Are there gender differences in the self–assessment of male and female surgical residents?	Male and female residents performed equivalently. All residents underestimated their abilities compared with faculty assessment; however, general surgery residents did so to a greater degree ($P < .05$). Female residents demonstrated a greater degree of underestimation than did their male colleagues; however, this was not statistically significant.	Although female resident surgeons are generally confident in their abilities, this may be in contrast to the self- perception of many female medical students. Consideration of gender differences in self-perception may be important when providing feedback to female students and residents.
19.	Paradise JE, Finkel MA, Beiser AS, Berenson AB, Greenberg DB, Winter MR (1997)	1.To measure agreement about genital examination findings among physicians who rate themselves as skilled in evaluating children for suspected sexual abuse 2.to compare these physicians' descriptions and	USA Physicians N=414	Questionnaires including seven simulated cases, each consisting of a brief history and one photograph of a girl's genitalia, were mailed to random samples of two groups: the members of four physician organizations concerned with child abuse or pediatric gynecology, and pediatricians at large. Among the surveyed physicians who rated	What degree of agreement was there between self rated experts in childhood sexual abuse and that of a panel of experts?	Responses were received from 548 (50.9%) of 1076 physicians; 414 responses (75.5%) were analyzable. Two hundred six physicians (50%) rated themselves as skilled in assessing children for sexual abuse. On average, 45% of these physicians' descriptions and 72.6% of their interpretations conformed with the consensus standards. In four cases, between 5% and 20.7% of these physicians described genital findings that the expert panel had considered absent from the photographs. Conformity with standard	Assessments of girls' genital findings by physicians who rate themselves as skilled in examining children for suspected sexual abuse often differ. In some cases, among physicians who all rate themselves as skilled, assessments made by very experienced physicians may conform more closely to consensus standards than do assessments made by less experienced physicians. Agreement with expert assessment depends on experience.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
		interpretations with consensus standards developed by an expert panel, 3. to investigate the effects of physician and case characteristics on agreement		their own skill in evaluating cases of suspected sexual abuse as higher than average, agreement was measured, both overall and between those with the most and with less clinical experience, and their conformity was asessed with consensus standard descriptions and interpretations.		interpretations tended to be higher in cases with photographs concordant with the accompanying, unambiguous histories (P=.06). The most experienced physicians resembled the expert panel more closely than did the less experienced self-rated skilled physicians in interpreting 3 simulated cases (P< or =.001).	
20.	Parker RW, Alford C, Passmore C (2004)	To determine if family medicine residents are able to self-assess their medical knowledge by predicting their performance on the In-training Examination (ITE).	USA Family medicine residents n = 311	Comparative Study Residents estimated their performance on the ITE in each of the 9 content areas just prior to undertaking the examination. Correlation coefficients were calculated for corresponding predicted and actual scores for each resident in each content area. Predictions were also compared to performance according to quartile.	1. How effectively can residents assess their knowledge, specifically by predicting their scores on the different topic areas in the ITE? 2. Whether knowledge of the subject matter was related to the ability to predict performance?	Residents' predicted performance did not correlate strongly with their actual performance. Pearson correlation coefficients for all residents in each content area and for the subcategories of men and women were all less than 0.3. The ability of men and women to predict their acores was similar. Added years of training did not improve the residents' ability to predict their performance. Residents scoring in either the lowest or highest quartile were least able to predict accurately with correct predictions ranging from 3% to 23%.	Residents cannot reliably predict their performance on the ITE. Of special concern are residents scoring in the lowest quartile since they greatly overestimated their performance.
21.	Reiter H, Eva K, Hatala R, Norman G (2002)	To test the effectiveness of the relative-ranking model in undergraduate problem-based learning (PBL) tutorials	Canada First year medical students n = 36	Comparative study. Students were provided with relative ranking forms listing seven domains of competence along with their definitions. The student, two of the student's peers	What will be the correlations between relative ranking of tutorial performance by self, peers and tutors?	No significant correlations were found, suggesting that the relative rankings assigned by study participants were not reliable (This held, even when data were re-analysed using only extreme traits). The average inter-rater reliability between self-assessments and peer	The relative rankings assigned by the participants were not reliable. The consistently low correlations suggest very strongly that the relative- ranking system is not a useful tool for allowing individuals to assess tutorial performance. The relative ranking model proposed by Regehr et al maintains promise as an evaluation tool that

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
				and the student's tutor completed the ranking exercise after their second, fourth and sixth tutorials.		or tutor assessments mimicked the poor correlations commonly found in the SA literature (self-peer assessments r=0.003, self-tutor assessments r=0.037). In contrast to past research the inter-rater reliability was also poor for tutor–peer assessments, r= -0.007).	can provide appropriate and timely formative feedback, it did not prove to be a reliable method of assessment in the tutorial context during this study.
22.	Rudy DW, Fejfar MC, Griffith CH, Wilson JF (2001)	To compare faculty, peer and self-assessment of interviewing skills during a first year communication and interviewing course	USA First year medical students n = 97	Comparative study Evaluation of student interview performance on video. Evaluation of same encounter was completed by all evaluators at same time. Differences between self, peer and faculty evaluation were examined using multiple regression in repeated measure framework contrasts between groups.	Do evaluations derived from the three evaluations (faculty, peer, and self-assessment) correlate highly with each other?	Correlations among peer, self and faculty composite ratings were moderate. Pearson 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 only showed marginal statistical significance (r=0.19, df=80,p=0.08). In terms of comments peer ratings were highest, followed by faculty and then self-ratings. Faculty supplied the largest number of positive comments, followed by peers and then comments by self. Individuals described themselves with the largest number of negative comments. Written evaluations showed peers were more lenient than faculty and students were most critical of their own performances.	Students were willing to provide positive as well as corrective written feedback regarding their peers. Peer assessment skills did not appear to translate to self-assessment skills. Student's were willing to evaluate their peers in a balanced fashion but their self-ratings were overwhelmingly critical. This finding is in line with similar studies which have taken place. First-year students are capable of evaluating their peers but have difficulty assessing their own performance. Further interventions are needed to foster self-assessment skills in first year medical students.
23.	Sommers PS, Muller JH, Ozer EM, Chu PW (2001)	To measure physician faculty members' perceived self-efficacy (SE) for performing key medical functions	USA Physician faculty members n = 54	Comparative Study. Pre and post programme measurement of mean self-efficacy scores for each of the nine professional role functions and a total self-efficacy score for all the functions combined.	How do four independent variables (faculty type, length of faculty experience, age and gender) affect self-efficacy scores?	1. Faculty type - teaching faculty group compared to the clinician educator group had a higher self-efficacy score and higher SE scores for each of the 9 professional role functions. 2. Time on faculty did not have a significant effect on the total SE score or on the scores for the 9 professional areas. (however all <3	Increasing length of time in a faculty position did not affect SE scores. Gender showed no statistically significant effect on SE score. Scores for different type of faculty raised questions on group specific faculty development needs.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
				Measurement was carried out by the self-efficacy instrument known as Faculty Self-efficacy Scale (FacSES).		years). Women had lower SE scores than men, but gender was not statistically significant on total SE scores or the nine areas. 4. Type of faculty was statistically significant in 2 of the professional areas (feedback to learner, career planning). 5. No statistically significant association was found between participant age and the total SE score or that for individual areas.	
24.	Sullivan ME, Hitchcock MA, Dunnington GL (1999)	To investigate the association between self, peer and faculty evaluations in the setting of a problem based tutorial group.	USA Third year medical students participating in surgical clerkship n = 154	Comparative study. Students were randomly assigned to problem-based learning groups and completed self and peer evaluations at the end of the last tutorial. The instrument used asked for ratings on a 5 point scale from poor to outstanding on the items of problem solving, independent learning and group participation. Students and peers used this same instrument for ratings. These evaluations were compared with expert tutor ratings using Pearson correlation coefficients.	1.Can students identify their own strengths and weaknesses as compared with their peers and faculty? 2.To what extent do peer and faculty ratings correlate in a problem based tutorial?	Highest correlation was found between peer and faculty ratings. There was a moderate correlation in independent learning (r=0.5) and group participation (r=0.54) and lower correlation in problem solving (r=0.24), all significant at p=0.001. Lowest correlation was found between self and faculty ratings. There was very low correlation in problem solving (r=0.11), a low correlation in independent learning (r=0.24) , significant at 0.01 level and a low correlation in group participation(r=0.18), significant at 0.05 level. In terms of self and peer ratings correlations were low but statistically significant: problem solving r=0.18, p=0.05); independent learning (r=0.21, p=0.01); group participation (r=0.23, p=0.01)	Students are not able to identify their own strengths and weaknesses as compared with their peers and faculty. There is only a moderate correlation between peer and expert ratings in a tutorial setting. Possible explanation for results is that students are not routinely taught self-evaluation skills in a traditional curriculum.
25.	Tracey J, Arroll B, Barham P, Richmond D (1997)	To determine whether general practitioners can make accurate self assessments	New Zealand General Practitioners n = 67	Comparative study. Random sample of General Practitioners completed a self assessment of their level	Can general practitioners make accurate self assessments of their knowledge?	Correlations between self assessments and test scores were poor for all three topics studied (r=0.19 for thyroid disorders, 0.21 for non-insulin dependent diabetes, 0.19	As general practitioners cannot accurately assess their own level of knowledge on a given topic, professional development programmes that rely on the doctors' self perceptions to assess their needs are likely

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
		of their knowledge in specific areas.		of knowledge over a variety of topics using a nine point semantic differential scale. An objective assessment of their knowledge was then made by administering true-false tests on two of the topics: thyroid disorders and non-insulin dependent diabetes. The study was repeated with another group of 60 general practitioners using sexually transmitted diseases as the topic		for sexually transmitted diseases).	to be seriously flawed.
26.	Ward M, MacRae H, Schlachta C, Mamazza J, Poulin E, Reznick R, Regehr G (2003)	To verify the accuracy of self-assessment for the performance of a surgical task, and to determine whether self-assessment may be improved through self-observation or exposure to relevant standards of performance.	Canada Senior surgical residents (n=27) and a general surgery fellow (n=1)	Comparative study. Comparison of self rating with gold standard rating of a calibrated rater, using a global rating scale (GRS) and the Operative Component Rating Scale (OCRS)	1. What is the accuracy of self-assessment for the performance of a laparoscopic operation? 2. Do interventions (self-observation of video-taped performance and review of benchmark performances) lead to an improvement in self-assessment ability?	There were three self-evaluations of performance: immediately after the performance of the surgical procedure; after self-observation of the videoed performance; after review of the four videotaped benchmark performances. The correlation between experts' evaluations and residents' initial self-evaluations was moderate (r=0.50, P<0.01). The correlation between experts' and self-assessments after self-observation of videotaped performance showed a statistically significant increase to 0,63 (D r=0.13, p0.01) suggesting that the opportunity to view one's own performance improved self-assessment accuracy. However the correlation between experts' assessments and residents' self-assessments after the residents had viewed the benchmark tapes was not significantly different, at r=0.66 (D r=0.03, not significant).	Senior surgical residents are fairly accurate judges of their technical performance in a laparoscopic model. Self-observation of videotaped performance improved the residents' ability to self-evaluate. However unlike other previous studies the opportunity to view benchmark performances of the same procedure did not improve self-assessment ability. This finding lends credence to the speculation that the ability to self-assess is related to (surgical) experience. Future investigation is required to determine whether the demonstrated self-assessment abilities are specific to the evaluation of technical performance or whether more senior trainees are better self-assessors in general.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
27.	Weiss PM, Koller CA, Hess LW, Wasser T (2005)	To investigate whether the third year medical students completing an obstetrics and gynecology clerkship assess themselves at the same level of clinical competence as do the residents/ faculty.	USA Medical Students n = 47	Comparative study. Penn State College of Medicine evaluation tool completed by residents/faculty about each of the students – covers 5 areas – fund of knowledge, personal attitudes, clinical problem solving skills, written/verbal skills, and technical skills. Students used an identical tool to rate themselves at the end of the clerkship, and also predicted their NBME shelf exam scores. Self assessed scores for the 5 skill areas were compared with the averaged score given by residents/ faculty.	How do medical students' self-assessments compare with their final clerkship grades and with their performance in the NBME Shelf Exam?	There was a statistically significant weak to moderate, positive correlation between students' self-assessment and final clerkship grade for written/verbal skills ($p = 0.002$, $r = 0.390$). 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 NBME score and categorized true score ($p = 0.717$, $r= 0.49$).	At the end of obstetrics and gynecology clerkship, third-year medical students are better at assessing their technical and written/verbal skills than their global fund of knowledge and personal attitudes. These results may suggest that students are not aware of their own personal attitudes and communication skills and how they can affect their effectiveness as a physician.
28.	Wilkerson L, Lee M, Hodgson CS (2002)	To evaluate the effects of an enhanced curriculum in cancer prevention on medical students' knowledge and self-perceived competency in the use of counseling and screening examinations	USA Medical students n=333 (144 baseline students, 95 first year students, 72 second year students, 52 third year students)	Comparative Study. Students completed a cancer prevention and detection survey. In addition they self rated their levels of competency in counselling for smoking prevention, smoking cessation, sun protection and healthy nutrition. They also rated their competencies in	What effect did the new curriculum have on student's knowledge and self-perceived competencies in counseling and screening skills?	**Knowledge**: the baseline students scored significantly ($p<0.001$) lower and the third year students scored significantly ($p<0.001$) higher than the other groups in the overall knowledge and most of the six sub-scales in the questionnaire. **Self-perceived counselling and screening skills**: the cohorts differed significantly ($p<0.001$) in their self-rated competencies with higher ratings associated with each higher level of training.	Findings of this study demonstrated support for the effects of an enhanced cancer prevention curriculum on medical students' knowledge, counselling and screening skills. In general the student's mean scores in these areas increased progressively for each cohort. In terms of the three educational strategies (practice, observation and direct instruction) relative to the three outcome measures, it is clear that for all sudents, practice is the major way in which curriculum affects self-perceived competency.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
		during each of the first three years of medical school. To evaluate how three instructional strategies used (direct instruction, hands-on practice and observation) contributed to these outcomes.		performing skin, breast, Pap smear and digital rectal screening examinations. Knowledge and self-perceived competencies in counselling and screening skills were the three outcome variables.		**Instructional methods**: The relative contributions to practice, observe or receive direct instruction were explored by multiple regression analysis. Combining all year groups the amount of self-reported practice was the single best predictor for all three outcome measures.	
29.	Woods R, McCarthy T, Barry MA, Mahon B (2004)	To assess knowledge, attitudes and practices in the diagnosis and management of vesicular rash illness	USA Primary care and emergency physicians n = 178	Comparative Study. Confidential paper survey questionnaire, eliciting data on perceived comfort in diagnosing and evaluating rashes, knowledge of key differential diagnostic characteristics of chickenpox and smallpox and diagnostic interpretation of colour photographs of patients with smallpox or chickenpox. Responses summarized as perceived comfort score, differential diagnosis score and picture score.	None stated beyond aim of study.	Most physicians felt comfortable evaluating a patient with a rash illness (69%) or in diagnosing chickenpox (89%) but few (17%) felt comfortable diagnosing smallpox. Those who were comfortable diagnosing rash illnesses had higher differential diagnosis scores. Experienced physicians had greater perceived comfort than junior physicians.	Strategies for bioterrorism-related training could take advantage of physicians' awareness of their own knowledge deficits.
30.	Woolliscroft JO, Tenhaken J, Smith J, Calhoun JG (1993)	To identify predictors of students' initial self-assessment of their clinical performances, the predictive value of their self-	USA Third year medical students n = 142 (137 participated)	Comparative Study. Students completed a self-assessment questionnaire at the beginning and end of their internal medicine clerkship. Questionnaires consisted of Likert scales	1) What are the factors that shape third year medical students' initial self-assessment? 2) What is the predictive value of third year medical	Weak to absent correlations were found between prior-performance measures and initial self-assessments. The lower performing students, as measured by college grade-point averages and Medical College Admission Test scores, tended to rate their performances	Findings are similar to previous studies in that they showed relatively poor agreement between external measures of students' performances and student's self-assessments of their performances.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
		assessments and factors that relate to their final self-assessments.		and students rated their abilities regarding clinical skills, use of knowledge in the clinical setting and discharge of patient care responsibilities. Data was also collected on student's performance as measured externally: college grade point averages; standard exams taken before, during and after the clerkship; and ratings given the students by the faculty and residents with whom they worked in the clerkship. Pearson product-moment correlations were then calculated between students' self-assessment ratings and their scores on the external measures of performance.	students' self-evaluations? 3) What factors are related to third year medical students' final assessments in a clerkship?	higher than did their peers at initial assessment. In contrast, the higher performing students rated themselves lower than would be warranted given their prior performances. There were significant increases in the initial self-assessments as the year progressed and the students entered the clerkship after having had more experience. The correlations between the students' final self-assessments and the ratings by faculty and residents were generally weak. The strongest (0.267, $p<0.001$) concerned the students' medical knowledge. In addition there was a moderate correlation (0.413, $p<0.001$) between the students' self-assessment of how hard they had worked and their self-assessments of overall performance.	
31.	Young JM, Glasziou P, Ward JE (2002)	To evaluate General Practitioners understanding of evidence based medicine terminology by self-rating using a blinded validation study.	Australia General Practitioners n = 50	Comparative study. Comparison of participants' self-rating of understanding of seven terms used in evidence based medicine with objective criteria developed by experts defining evidence based medicine terms. Used self-administered questionnaire and structured interview to test understanding.	Do GPs understand evidence based medicine terminology?	Positive and negative predictive values were calculated for each term to assess the probability of competence given a positive or negative self-rating. The predictive value of a positive self-rating was 8% for the term "positive predictive value" but zero for the six other terms. No participants demonstrated competence exceeding their self-rating meaning the predictive value of a negative self-rating was 100% for all terms.	Participants' self-rating of their understanding of terms used in evidence based medicine differed from an objective, criterion based assessment. Study concludes that GPs do not understand the principles of evidence based medicine.

	Authors	Objectives	Setting, Population and Numbers	Type of Study Study Design	Research Questions	Results	Conclusion
32.	Zonia SL, Stommel M (2000)	To compare the evaluations that interns made of themselves with those of their faculty.	USA Osteopathic Interns n = 73	Comparative Study. Examined paired differences in self and trainer ratings and also differences between different groups of interns (fast trackers and traditional interns). Using an instrument with ordinal item response (1=unacceptable to 4 =very good), all of the interns evaluated themselves in two major dimensions (academic growth and professional development). Faculty used the same instrument to rate the interns. The interns evaluated themsleves at the end of each rotation, before seeing their faculty-trainers' evaluations	How do interns' self - ratings compare with those of their faculty?	The fast trackers rated themselves consistently lower than the traditional interns especially with regard to their medical knowledge and skills (mean ratings of 3.18 versus 3.62; p=0.001) but also in terms of their professional relations and personal growth (3.63 versus 3.87; p=0.001). The faculty also rated fast trackers somewhat lower than they rated the traditional interns. The interns, regardless of type, consistently rated themselves lower than did their faculty trainers (3.63 versus 3.86; p=0.001). The sex of the intern had no bearing on the ratings, whether self or faculty ratings. Finally as the interns progressed through their rotations , both their self-ratings and their faculty's ratings consistently increased through the fifth month (p=0.001). After the fifth month the ratings reached a plateau.	Faculty viewed their trainees more positively than did the trainees themselves. Fast track interns were judged less favourably by themselves and to a lesser extent by the faculty. The women accepted into the programme were no more or less critical of their performance than were their male colleagues, and the faculty were no more or less critical of female or male trainees. The data supports the hypothesis that self-ratings will mirror those of trainers over time.

Table 2 Excluded Papers and Reason for Exclusion

Authors	Year	Journal	Title	Reason for Exclusion
Amery J, Lapwood S	2004	Palliative Medicine	A study into the educational needs of children's hospice doctors: a descriptive quantitative and qualitative survey.	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment.
Arthur H	1995	International Journal of Nursing Studies	Student self-evaluations: How useful? How valid?	Review paper only - no original research
Asch E, Saltzberg D, Kaiser S	1998	Academic Medicine	Reinforcement of self-directed learning and the development of professional attitudes through peer-and self-assessment.	Poor reporting/insufficient information
Belar C, Brown RA, Hersch LE, Hornyak LM, Rozensky RH, Sheridan EP, Brown RT, Reed GW	2001	Professional Psychology: Research and practice	Self-assessment in Clinical Health Psychology: a model for ethical expansion of practice	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Biran LA	1991	Medical Education	Self-assessment and learning through GOSCE group objective structured clinical examination	Group assessment rather than self-assessment
Butterfield B, Metcalfe J	2001	Journal of Experimental Psychology	Errors committed with high confidence are hypercorrected	Outwith clinical context
Crawford MW, Kiger AM	1998	Journal of Advanced Nursing	Development through self-assessment: strategies used during clinical nursing placements	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Dornan T, Boshuizen H, Cordingley L, Hider S, Hadfield J, Scherpbier A	2004	Medical Education	Evaluation of self-directed clinical education: validation of an instrument	Not about self-assessment
Durieux P, Bissery A, Dubois S, Gasquet I, Coste J	2004	Quality and Safety in Health Care	Comparison of health care professionals self-assessments of standards of care and patients opinions on the care they received in hospital: observational study	Group assessment rather than self-assessment
Edwards A, Robling M, Matthews S, Houston H, Wilkinson C	1998	British Medical Journal	The vast range of clinical conditions means that doctors cannot know everything	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Ehrlinger J, Dunning D	2003	Journal of Personality and Social Psychology	How Chronic Self-Views Influence and Potentially Mislead Estimates of Performance	Outwith clinical context

Authors	Year	Journal	Title	Reason for Exclusion
Evans AW, Leeson R, Newton-John TRO	2002	Medical Education - Really Good Stuff	Influence of self-deception and impression management on surgeons' self-assessment scores	Poor reporting/insufficient information
Garland G	1996	Journal of Nursing Staff Development	Self Report of Competence: a Tool for the Staff Development Specialist	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Gordon MJ	1997	Academic Medicine	Cutting the Gordian Knot: A Two-part Approach to the Evaluation and Professional Development of Residents	Review paper only - no original research
Henderson P, Johnson M	2002	BMC Medical Education	An innovative approach to developing the reflective skills of medical students	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Indulski JA	1999	International Journal of Occupational Medicine and Environmental Health	Self-assessment of competence in public health management as a measure of effectiveness of training	Self-assessment used as a blind tool
Jacob J, Ostechega Y, Gallaway L	1990	Cancer Nursing	Self-assessed learning needs of oncology nurses caring for individuals with HIV-related disorders	Self-assessment used as a blind tool
Khan KS, Davies DA, Gupta JK	2001	Medical Teacher	Formative self-assessment using multiple true false questions on the Internet: feedback according to confidence about correct knowledge	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Lofgren M	1996	Medical Education	Self-marking in Written Examination: A Way of Feedback and Learning	Not about self-assessment
Malkin KF	1994	Journal of Nursing Management	A standard for professional development: the use of self and peer review; learning contracts and reflection in clinical practice	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment

Authors	Year	Journal	Title	Reason for Exclusion
Manogue M, Brown GA, Nattress BR, Fox K	1999	International Endodontic Journal	Improving student learning in root canal treatment	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
McCord EC, Smorowski-Garcia K, Doughty A	1997	Academic Medicine	Assessment at one school of students abilities and confidence in diabetic patients education	Not about self-assessment
Mires GJ, Friedman Ben-David M, Preece PE, Smith B	2001	Medical Teacher	Educational benefits of student self-marking of short-answer questions	Not about self-assessment
Mussweiler T, Strack F	2000	Journal of Personality and Social Psychology	The Relative Self: informational and judgmental consequences of comparative self-evaluation	Outwith clinical context
Norcini J, Lipner R, Downing SM	1996	Academic Medicine	How meaningful are scores on a take home recertification examination?	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Norman GR, Shannon SI, Marrin ML	2004	British Medical Journal	The need for needs assessment in continuing medical education	Review paper only - no original research
Reisine S	1996	Journal of Dental Education	An overview of self-reported outcome assessment in dental research	Not about self-assessment
Schmidli-Bless C	1999	Pflege	Quality assurance in nursing: self evaluation and peer review of nursing standards. Review of 2 years experience	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Schwiebert LP, Davis A.	1995	Teaching and Learning in Medicine	Impact of a Required Third-Year Family Medicine Clerkship on Student Confidence in Cognitive and Procedural Skills	Self-assessment used as a blind tool
Seidenberg M, Haltiner A, Taylor MA, Hermann BB, Wyler A.	1994	Journal of Clinical and Experimental Neuropsychology	Development and validation of a multiple ability self-report questionnaire	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Sharp LK, Wang R, Lipsky MS	2003	Academic Medicine	Perception of Competency to Perform Procedures and Future Practice Intent: A National Survey of Family Practice Residents	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Sobral DT	2000	Medical Education	An Appraisal of Medical Students' Reflection in Learning	Not about self-assessment

Authors	Year	Journal	Title	Reason for Exclusion
Sobral DT	2004	Medical Teacher	Medical students' self-appraisal of first -year learning outcomes: use of the course valuing inventory	Not about self-assessment
Steucheli S, Hand R	1993	Medical Education	Participants' perceptions versus actuality	Poor reporting/insufficient information
Van Rosendaal GMA, Jennett PA	1992	Academic Medicine	Resistance to peer evaluation in an internal medicine residency	Group assessment rather than self-assessment
Vecchioli A, Ferro-Luzzi M, Campioni P	1990	Rays - Rome	Assessment and self-assessment	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Wakley G	2000	Sexual and Relationship Therapy	Sexual health in the primary care consultation: using self-rating as an aid to identifying training needs for general practitioners	No intervention, evaluation of self-assessment or information about attitudes towards self-assessment
Westberg JH	1994	Family Medicine	Fostering learners' reflection and self assessment	Review paper only - no original research