Advances in Health Sciences Education **7:** 63–80, 2002.

Reflections

Measuring Self-assessment: Current State of the Art

MYLÈNE WARD[1], LARRY GRUPPEN[2] and GLENN REGEHR[3*]
[1]*Department of Surgery, University of Toronto, Toronto, Ontario, Canada;* [2]*Department of Medical Education, University of Michigan, Ann Arbor, Michigan, USA;* [3]*Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada (*author for correspondence; Centre for Research in Education, University Health Network, 200 Elizabeth St., 1 Eaton South, Toronto, Ontario M5G 2C4, Canada. E-mail: g.regehr@utoronto.ca)*

Abstract. The competent physician pursues lifelong learning through the recognition of deficiencies and the formulation of appropriate learning goals. Despite the accepted theoretical value of self-assessment, studies have consistently shown that the accuracy of self-assessment is poor. This paper examines the methodological issues that plague the measurement of self-assessment ability and presents several strategies that address these methodological problems within the current paradigm. In addition, the article proposes an alternative conceptualization of self-assessment and describes its associated methods. The conclusions of prior research in this domain must be re-examined in light of the common pitfalls encountered in the design of the studies and the analyses of the data. Future efforts to elucidate self-assessment phenomena need to consider the implications of this review.

Key words: education, educational measurement, medical, self-assessment, self-evaluation programs

Introduction

As with any profession that operates under the principles of self-regulation and autonomy, the competent physician must be a self-directed, lifelong learner (Moore and Cordes, 1992). The first step in this process is the diagnosis of one's own learning needs, which enables the formulation of appropriate learning goals (Spencer, 1999). Therefore, the ability to accurately assess one's strengths and weaknesses is critical to the enterprise of lifelong learning (Gordon, 1992). Despite this theoretical argument for the critical importance of self-assessment in the professions, the conclusions drawn from research regarding professionals' ability to self-assess are mixed at best.

Two previous reviews offer a comprehensive summary of findings in the self-assessment literature. In 1989, Falchikov and Boud published a meta-analysis of quantitative self-assessment studies in higher education. They reported the results of forty-four studies in a variety of subject areas, including medicine, guidance counseling, law, engineering, behavioral science, psychology, and dietetics. Correlations between self-assessed and external measures of performance ranged from

–0.05 to 0.82, with a mean correlation of 0.39, suggesting that, on average, self-assessors were poor to moderate judges of their performance. Subsequently, Gordon (1991) published a literature review of eighteen self-assessment studies in the health professions. Studies of self-assessment of factual knowledge reported correlations in the range of 0.02 to 0.65. The accuracy of global self-assessments based on an extended period of performance was even worse, with the highest correlation reported among the six studies in this category being 0.32. Both reviews suggest that the measurement of self-assessment often yields less than promising results.

Since self-assessment is fundamental to the concept of self-directed learning and the maintenance of professional competence, educators find it troubling that researchers who have attempted to establish the accuracy of self-assessment have often observed incongruities between self-evaluations and external measures of achievement. But researchers' conclusions about the accuracy of self-assessment have been tempered by multiple considerations. In particular, the measurement of self-assessment encounters several methodological problems. This may limit the literature's capacity to support conclusions about self-assessment ability. Our objective in this paper is to examine the quality of the measurement of self-assessment. We begin by providing a synopsis of the methodologies employed in the self-assessment literature. We then examine the problems inherent in the predominant methodological approach. In the last section, we present several strategies that address these methodological issues. In addition, we propose an alternative conceptualization of self-assessment and describe its associated methods.

Summary of the Literature: Methodologies

In order to investigate approaches to the measurement of self-assessment, we conducted a review of the methods employed in the self-assessment literature. We included 41 of the studies reviewed by Falchikov and Boud (1989), eliminating 3 unpublished studies. Of the 18 studies identified by Gordon (1991), six appeared in the Falchikov and Boud (1989) meta-analysis, three did not provide a quantitative measure of self-assessment accuracy, and three studies were excluded since they could not be located, leaving a total of 6 unique articles. Additional studies published since 1990 were identified by a computerized literature search of the MEDLINE, CINAHL, Education Resources Information Center, and PSYCHINFO databases. These searches used combinations of the key words: *self-assessment*, *self-evaluation*, *self-concept*, *higher education*, and *medical education*. Bibliographies were searched for relevant articles. We included all studies if they met the following criteria: a) a study population of individuals in higher education and, b) a quantitative measurement of self-assessment accuracy.

A total of 67 studies (41 from Falchikov and Boud, 6 from Gordon, and 20 new studies) were included in this analysis. A review of the methodologies employed

in each indicates that the measurement of self-assessment accuracy is approached in a similar fashion across a range of disciplines in higher education and the health professions. Table I presents a classification of studies by method.

The most common methodology (41 studies) used to evaluate self-assessment involves correlational analyses. In this common design, a self-assessment score and a score based on some external measure (often an expert evaluation) is generated for each individual in the group. Across the group the self-ratings are correlated with the expert ratings to obtain a single numerical value for the group. This numerical value is interpreted as a measure of the group's self-assessment ability. The assumption is that if all students are effectively evaluating their ability relative to their peers, the collection of student self-assessments should correlate well with the external measure. Conversely, if the correlation is low, it is assumed that the group as a whole is poor in determining the quality of their performance relative to their peers.

As a separate but related methodology, some studies (N = 16) reported the proportion of self-ratings that corresponded with expert ratings. Falchikov and Boud (1989) observed that the definition of 'agreement' is not consistent across this group of studies. Some require identical ratings on a 100 point scale, but most studies adopt a slightly more liberal definition of 'agreement', whereby self-ratings are deemed equivalent to expert ratings if they assign matching scores on a Likert-type scale. Both 'percentage agreement' and 'correlation' are techniques that are used to describe the relationship between two variables, in this case self and expert assessment. Both approaches generate conclusions for the group with respect to self-assessment ability, as defined by close agreement between scores generated by trainees and experts. Thus, while somewhat different in presentation, the correlational and agreement designs share many of the methodological issues that we will be discussing. A total of 55 studies (82% of studies identified in the review) employed one or both of these methods.

A third methodological approach that appears frequently in the self-assessment literature (27 studies) involves the direct comparison of the absolute values of self-ratings and some external standard. This approach provides information about whether student self-ratings tend to match the external standard, or whether the self-assessors tend to overrate or underrate themselves. As with the other methodologies, the intent generally involves comparing the self-assessing group as a whole to the external standard by comparing means.

It is predominantly on the basis of these three quantitative methodologies that claims about self-assessment ability have been made. With the exception of a small number of articles that we have published recently, every paper making quantitative evaluations of self-assessment used one or more these techniques exclusively. Yet there has been relatively little reflection on the nature of these methodological paradigms and the assumptions under which they function. The following section will describe some of the methodological pitfalls that face these models for evaluating self-assessment ability.

Table I. Summary of methodologies employed in quantitative studies of self-assessment*

Field	Correlation	Percentage agreement	Group mean comparisons	Inter-individual approach
Health Professions	Antonelli 1997 Arnold 1985 Calhoun 1988 Daniel 1990 Das 1998 Everett 1983 Farnill 1997 Hay 1995 Henbest 1985 Herbert 1990 Johnson 1998 Kaiser 1995 Kolm 1987 Leichner 1980 Linn 1975 MacFadyen 1985 Martin 1998 Morton 1977 Palmer 1985 Plorde 1985 Rezler 1989 Risucci 1989 Stuart 1980 Wooliscroft 1993	Cochran 1980 Coutts 1999 Forehand 1982 Henbest 1985 Kaiser 1995 Mast 1978 Sclabassi 1984	Calhoun 1988 Calhoun 1990 Daniel 1990 Das 1998 Farnill 1997 Geissler 1973 Hay 1995 Henbest 1985 Herbert 1990 Johnson 1998 Morton 1977 Palmer 1985 Risucci 1989 Stuart 1980 Zonia 2000	Fitzgerald 2000 Gruppen 1997 Gruppen 1998 Harrington 1997 Regehr 1996
Subtotal 37 studies	24 (64.9%)	7 (18.9%)	15 (40.5%)	5 (13.5%)
Higher Education	Bergee 1997 Bishop 1971 Boud 1979 Boud 1986 D'Augelli 1973 Doleys 1963 Gaier 1961 Irvine 1983 Israelite 1983 Keefer 1971 LeBlance 1985 Mihal 1984 Murstein 1965 O'Neill 1985 Pease 1975 Pohlmann 1974 Wheeler 1981	Burke 1969 Davis 1980 Falchikov 1986 Filene 1969 Gray 1987 Mueller 1970 Pitishkin-Potanich 1983 Stanton 1978 Stover 1976	Bishop 1971 Boud 1986 Chiu 1975 Doleys 1963 Fuqua 1984 Greenfield 1978 Israelite 1983 Keefer 1971 McGeever 1978 Mihal 1984 O'Neill 1985 Wheeler 1981	
Subtotal 30 studies	17 (56.7%)	9 (30.0%)	12 (40.0%)	–
TOTAL 62 studies (excluding 'Intraindividual')	41 studies (66.1%)	16 studies (25.8%)	27 studies (43.5%)	–

*Listed by first author only for ease of presentation.

Methodological Issues

This discussion begins with a focus on the predominant approach to the measurement of self-assessment, the correlational model. As elaborated above, correlational analyses are used to determine whether a group of individuals are accurate judges of their own performance, as compared to an external measure of performance. A weak or absent relationship between self-ratings and the external measure suggests that the individuals in this group are poor self-assessors. We have identified three methodological issues with the correlational design that cast doubt on this conclusion: problems with the gold standard, problems with differential use of the scale by participants, and problems with group level analyses.

Problems with the 'gold standard'

First, it is worth noting that expert (e.g. medical faculty, clinical preceptors) evaluations are a common source of the 'objective' measure against which student self-assessments are compared. For evaluation of a specific task, experts are chosen to observe and judge the performance. For the evaluation of performance during a clinical rotation, preceptors are asked to provide a global assessment. Thus, all studies that compare self-assessment to expert assessment and draw a conclusion about self-assessment ability are making the same basic assumption: expert judgement is the gold standard by which to measure all aspects of clinical competence. In other words, faculty and supervisors provide the 'true rating' (Palmer et al., 1985).

As many authors recognize, the *reliability* claims of this gold standard are suspect (Abrams and Kelly, 1974; Arnold et al., 1985; Bergee, 1997; Falchikov and Boud, 1989; Farnill et al., 1997; Harrington et al., 1997; Hay, 1995; Johnson and Cujec, 1998; Kolm and Verhulst, 1987; Martin et al., 1998; Regehr et al., 1996). However, only a handful of studies report the reliability of the gold standard, and for these few studies, there is evidence of inconsistency among expert raters. Harrington et al. (1997) applied a relative ranking model to the global assessment of performance over an orthopedic residency rotation and reported a mean inter-rater reliability of only 0.27. The authors hypothesized that study designs using longitudinal evaluations are particularly prone to rater unreliability. Over the course of a residency or clerkship rotation, each preceptor observes trainees in different clinical settings (e.g. at rounds, in the OR, on the ward) performing different clinical responsibilities. Evaluations take place at the end of a lengthy rotation and are thus limited by recall.

By contrast, expert raters are far more likely to agree given the chance to evaluate a short, structured, and relatively simple task. The study by Regehr and colleagues (1996) lends support to this argument. The relative ranking model was applied to the assessment of a standardized patient interview. The mean inter-rater reliability was much higher, at 0.70. Martin et al. (1998) also reported high inter-reliability between two communication experts (Cronbach's alpha = 0.94) who viewed the videotapes of residents performing a standardized patient interview. The

question of expert reliability is not limited to medical education, as Bergee (1997) reported mixed inter-rater reliabilities (coefficient alpha 0.23 to 0.93) for evaluations of applied music performances. Thus, the idea of an infallible 'expert rater', the traditional gold standard, should be viewed with a degree of skepticism and this places important limitations on the interpretation of low correlations between expert and self ratings.

Furthermore, the *validity* of the 'gold standard' in these studies may be similarly questioned. For the purposes of self-assessment evaluation, the gold standard scores must satisfy two related conditions regarding validity. First, experts must provide a valid measure of the dimension that they claim to evaluate. This issue is addressed in an extensive body of literature on global rating scales (Gray, 1996; Keynan et al., 1987; Turnbull et al., 1998), however as just one example from the self-assessment literature, Risucci and colleagues (1989) reported that the mean supervisor ratings of cognitive achievement correlated only moderately with the total raw score on the in-training examination of the American Board of Surgery (ABSITE) ($r = 0.55$, $p < 0.01$). In this study, cognitive qualities accounted for 30% of the variance in faculty ratings, leaving 70% of the variance unexplained. With respect to the evaluation of the non-cognitive areas of competence, such as professionalism and communication skills, there are no real 'gold standards' (Gray, 1996; Keynan et al., 1987; Turnbull et al., 1998). Therefore, the validity of expert assessment remains elusive and uncertain.

However, even if this first validity concern is addressed, the validity of the expert rater presupposes a second condition: that the expert's concept of the dimensions that are relevant to a 'good' performance are reasonable and appropriate. For example, several studies suggest that clinical supervisor assessments reflect an emphasis on cognitive achievement while medical students emphasize non-cognitive abilities (Arnold et al., 1985; Kegel-Flom, 1975). Under these circumstances, a measure of the relationship between self-ratings and supervisor ratings would likely generate a low correlation, which, typically, implies inaccurate self-assessment. Arnold and colleagues (1985) argue that 'the validity of self-evaluations can not be established solely against criterion measures of cognitive achievement or faculty assessments that contain a cognitive emphasis'. In other words, even if one assumes that clinical supervisors are in fact accurate (valid) judges of cognitive achievement, any conclusion with respect to the accuracy of self-assessment presumes that experts are providing a fair measure of clinical performance by focusing on cognitive achievement.

In summary, it appears that there are significant problems with the assumption that experts' assessments qualify as the gold standard against which to judge self-assessment. Few efforts have been made to study the reliability of this gold standard, when this reliability functions as a theoretical upper limit on the correlation of the students with the gold standard. In addition, the validity claims of the gold standard must be further examined. There is good reason to believe that

experts are not measuring what they intend to measure, or that what they intend to measure is necessarily what is really important.

PROBLEMS WITH DIFFERENTIAL USE OF THE SCALE AMONG STUDENTS

The correlational approach to evaluating self-assessment also rests on the notion that it is appropriate to consider a group of individual self-assessment scores as a set of coherent scores. This notion makes two assumptions. If either of these assumptions is incorrect, the paradigm is severely limited in its ability to produce a true measure of self-assessment ability in the population being assessed.

First, in order to treat the individual self-assessments as a set of coherent scores, we must assume that the group of trainees are all evaluating themselves by tapping into the same aspect of performance. As mentioned earlier, differences *between* rater types (self, expert) have been explored, but no one has asked whether or not differences exist *within* the group of students. It seems improbable that students or residents constitute a monolithic group, that self-assessments would reflect a common understanding of the dimensions of performance. Thus, the finding that students tend to focus on non-cognitive aspects of performance (Arnold et al., 1985; Risucci et al., 1989) should not be interpreted as a claim that all students do so. Such an interpretation likely over-generalizes the consistency among students. If so, each individual student might be accurately evaluating the dimensions that he or she chooses to evaluate, but there is no reason to believe that the numbers generated from such accurate self-assessments would have any coherent collective meaning.

However, it would not be sufficient to show that all individuals are measuring themselves based on the same criteria. This research design assumes that all individuals measure these dimensions of competence in a consistent manner, and yet, even the best scale is subject to interpretation. As an extreme case, consider the following theoretical example (Table II). Four students (A, B, C, D) evaluated both themselves and each other member of the group on a 7 point scale. An expert used an identical instrument to evaluate the students. This table shows that each student gave his or her own performance a 5 out of 7. Thus, if we look only at the students' self-assessment scores (ignoring, as the paradigm usually does, the scores that the students would give to their colleagues), only student B is a good judge of her own performance, relative to the expert's assessment. Student A underestimated her performance and Students C and D overestimated their performances. On the surface, the group as a whole is self-assessing very poorly. Since all four students gave themselves 5 out of 7, the correlation of the students' self-assessments with the expert's assessments is zero. However, a closer look at the self-evaluations in the context of the students' peer evaluations shows that, in fact, self and peer assessments were all perfectly accurate (assuming the experts' assessments qualify as the gold standard). Student A, rated the highest by the experts, did recognize her performance was superior to her peers'; however she failed to use the upper-

Table II. Self, Peer, and Expert Assessments: an example to illustrate the impact of differential use of the scale.

Student	Expert	S_A*	S_B*	S_C*	S_D*	Self-assessments
A	7	5†	7	7	6.5	5
B	5	4	5†	6	6.0	5
C	3	3	3	5†	5.5	5
D	1	2	1	4	5†	5

*S_A denotes scores generated by Student A, S_B denotes scores by Student B, etc.
†Self-assessment scores.

most values of the scale. Student D, using a very narrow range of the scale, also demonstrated an understanding that his performance was poor relative to the group. This example illustrates four different interpretations of a score of 5 out of 7. It could be considered superior, average, or poor, depending on the individual and the context. This example also shows that inconsistent use of the scale among students attenuates the correlation of expert and self-ratings, regardless of the group's self-assessment ability.

By themselves, numbers are meaningless. The way in which each rater (self and expert) interprets and applies the scale must be defined before the numbers can be meaningfully combined and conclusions can be drawn about the accuracy of self-assessment.

PROBLEMS WITH GROUP-LEVEL ANALYSES

Even assuming that 'experts' are reliable and valid assessors of trainee performance and that students are measuring the same dimensions of performance using the scale in the same way, a conclusion about the accuracy of self-assessment based on the group correlation remains potentially problematic. This prominent methodological approach rests on yet another assumption: that every individual in the group is equal in terms of self-assessment ability.

Group-level correlation between self and expert ratings can make claims only at the level of the group. Either the correlation is low, suggesting that the group as a whole cannot self-assess effectively, or the correlation is high, suggesting that the group, as a whole, can self-assess accurately. Of course, the question of what is an acceptable correlation as a measure of self-assessment ability is likely to be a matter of further controversy. However, whether self-assessment is determined to be poor, moderate or good, the determination of this fact with a single correlation is of limited value. Further analyses and interpretation of a potentially complex phenomenon are difficult at best when we cannot understand individual variation within the group.

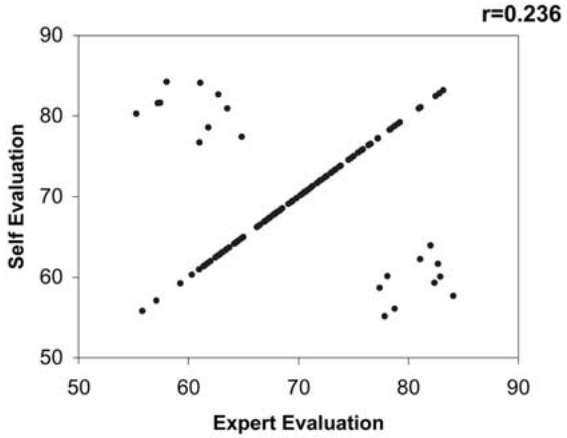

Figure 1. Self vs. Expert Evaluation: an example to illustrate the impact of a few 'poor self-assessors' on the overall correlation of self and expert evaluation scores.

Not only is individual variation within the group masked in this methodological paradigm, variation may also have a significant negative impact on overall group correlation. A few outliers could easily spoil the measurement of self-assessment ability for the whole group. The following hypothetical example illustrates this problem (Figure 1). The scatterplot shows that most individuals in this group (100 of 120) made accurate self-assessments of their performance. The correlation of these 100 scores with the gold standard is perfect, r = 1.0. However, ten individuals were above average but grossly underestimated their performances, and conversely, ten individuals were below average yet overestimated their performances. The correlation between students and experts for the whole group was only 0.236. These few outliers exercised a hugely negative impact on the overall correlation, even though the vast majority of individuals in the group were 'perfect' self-assessors. Although this example has been made extreme to demonstrate the point, it is nonetheless the case that group-level analyses fail to reflect an appreciation for group heterogeneity and therefore may skew the conclusions inappropriately.

PROBLEMS WITH THE DIRECT COMPARISON OF THE SCORES

Almost 50% of the studies reviewed make a direct comparison of student self-ratings to scores from some external criterion instead of or in addition to using the correlational method. This form of comparison is subject to many of the same methodological criticisms that are inherent in the correlational paradigm. Certainly, the assumption that the gold standard is a reliable and valid measure of the individual remains unsubstantiated. This model also assumes that all individuals are evaluating the same dimensions and using the scale in the same manner as the expert (as demonstrated in Table II). Further, the comparison of group means as generated by the participants and some external standard clearly hides individual

differences (every member of the group may be wildly inaccurate but the group mean could be identical to the mean of the external standard).

It is worth noting that two studies that directly compare scores draw conclusions about subgroups, often discovering that the high achievers tend to underestimate their performance relative to the gold standard, and underachievers tend to overestimate their performance relative to the gold standard (Arnold et al., 1985; Wooliscroft et al., 1993). These findings, however, may simply be a restatement of the fact that there is a weak correlation between self-assessment scores and scores generated by the gold standard. In statistical terms, a correlation of less than 1 necessarily implies regression to the mean, such that higher scores tend downward and lower scores tend upward. There sometimes appears to be a suggestion that overestimation by underachievers and underestimation by overachievers has motivational or ego defensive roots (e.g. Gordon, 1992). It is important to remember, however, that the same phenomenon would occur if a random number generator produced the self-assessment scores.

Dealing with the Pitfalls

We have described several major methodological issues associated with the traditional approach to the study of self-assessment. We believe that these problems are sufficiently severe to raise questions regarding the validity of the paradigm as it is usually instantiated, and to raise questions regarding the legitimacy of the general conclusions that might be drawn from the literature. It may be that some of these problems are insurmountable. However, several can be ameliorated, if not eliminated. In this section, we present several strategies for dealing with at least some of the 'pitfalls' that we have identified above. These strategies fall into two general categories. The first involves maintaining the existing paradigm and making alterations that minimize some of its obvious methodological weaknesses. The second involves the presentation of an alternative paradigm that we have been exploring for several years.

Solutions within the paradigm

'Solutions within the paradigm' refers to strategies that tackle the issues outlined in the previous section: Is our gold standard infallible and how would we know? Are students uniform in their evaluation of performance and their use of the scale? Is group-level analysis an oversimplification of self-assessment phenomena?

Correcting for the unreliability of expert ratings

It was suggested earlier that 'expert' evaluations of students might be neither valid nor reliable measures against which to compare student self-evaluations. Although we have no claims to more valid alternatives to evaluations provided by experts (there are no true gold standards in educational evaluation), it is relatively easy to

address issues regarding the reliability of these expert evaluations. The best way to improve the gold standard is to optimize reliability through the use of multiple expert raters. Agreement between raters will never be perfect, however expert unreliability can be taken into account when subsequently calculating the correlation between self and expert ratings. Regehr and colleagues (1996), for example, describe the use of the 'correction for attenuation' formula, whereby the student-expert correlation is divided by the square root of the expert inter-rater reliability. In this particular study, correlation between students' and experts' scores was 0.43. But when the student/expert correlations were corrected for the attenuation due to experts' unreliability, the mean self-assessment score increased to 0.58.

Stabilizing students' use of the scale

As with experts, we questioned the consistency with which the target group of students was evaluating the same criteria and using the scale consistently. Perhaps the best solution to the possibility that students are evaluating different dimensions of performance was enacted by Henbest and Fehrsen (1985), who asked medical students to create their own criteria for evaluation at the beginning of a family medicine rotation. When these same criteria were used to evaluate the students at the end of the rotation, a high positive correlation was found between self and faculty ratings (0.74; $p < 0.01$).

Problems with differential use of the scale among students can be addressed through the provision of explicit anchors for the evaluation criteria. Martin et al. (1998), for example, attempted to resolve this methodological issue by providing residents with exposure to 'benchmarks' of performance. Family practice residents performed an interview with a standardized patient and rated their performance. They were then shown videotapes of four performances of the same scenario that varied in quality and were asked to rate the four performances. To account for inconsistent use of the scale, the resident's self-assessment scores were *rescaled* based on their evaluations of the videotaped performances. The mean and standard deviation of each resident's scores for the videotapes were calculated, and the resident's self-evaluation was expressed as a z score relative to this mean and standard deviation. Thus the scale was standardized across residents through the calculation of z scores, an expression of where each resident felt they stood relative to the four performances. In this study, rescaling the residents' self-assessments into z scores did not improve the resident-expert assessment correlations (although, see Hodges et al., 2001, for a reanalysis of these data). However, this should not deter future efforts to apply these strategies or develop new innovative ways to ensure consistent use of the scale.

It may not be necessary, however, to use such an elaborate method for rescaling students' scores to ensure consistency in their use of the scale. Another possible strategy is to provide examples of videotaped performances with predetermined scores. These explicit anchors would likely help to standardize the scale.

Comparing self-assessment to peer assessment

It is worth acknowledging that the study of peer assessment has traditionally employed the same methodological paradigm as we have been describing and, thus, falls prey to the same potential pitfalls. Yet, several studies show that peer assessment is more accurate than self-assessment (Bergee, 1997; Falchikov and Goldfinch, 2000; Linn et al., 1976; Martin et al., 1998; Morton and MacBeth, 1977; Risucci et al., 1989). These studies lend support to the argument that individuals can identify good and bad performances, but are unable or unwilling to apply the same standards to their own performance. This use of peer assessment as a 'control' condition to evaluate self assessment may be an interesting alternative method to deal with the pitfalls that we have been describing. Rather than eliminating or ameliorating the pitfalls, we might control for them instead. Of course, correlations are notoriously unstable and demonstrating difference between two correlations (self assessment vs. peer assessment) may be difficult and may require more power than most studies can feasibly produce with the limited number of subjects available. In principle, however, this is certainly a useful option when the power is available.

AN INTRAINDIVIDUAL APPROACH

Rather than attempt to deal with the problems inherent to the 'traditional' paradigm, it is also possible to step outside the paradigm. We can search for new ways to conceptualize self-assessment and develop methods associated with this reconceptualization.

The development of a new framework for the study of self-assessment has been the recent focus of research in this domain. Regehr et al. (1996), for example, offered a reconceptualization of self-assessment that focused not on the individual's ability to rate herself relative to her peers, but on her ability to identify her own strengths and weaknesses relative to each other. They suggested that the ability to identify areas of performance that require the greatest degree of improvement would lend greater efficiency to self-directed learning efforts. Consistent with this framework, Gruppen and colleagues (1997, 2000; Fitzgerald et al., 2000) make a distinction between the conception of self-assessment as an *inter*individual process or as an *intra*individual process. The study of self-assessment according to the common methodological paradigm is an interindividual process. The subject asks, 'How good am I?' and generates a self-rating based on his perception of his ability relative to others or to some ideal standard. The intraindividual process is more consistent with the model suggested by Regehr et al. (1996) where self-assessment is defined as the ability of each individual to identify his or her own relative strengths and weaknesses. In other words, the self-assessor asks the question 'What aspects of my performance need the most work? Which aspects need the least work?'

Methodologically, this new intraindividual perspective requires multiple self-assessments from each subject in order to calculate individualized estimates of self-assessment accuracy. This is accomplished through the evaluation of multiple tasks or multiple aspects of a single performance.

An example of evaluating intraindividual self-assessment across a set of related tasks is provided in a study by Fitzgerald et al. (2000). Medical students completed an OSCE-type examination with ten stations (e.g. breast examination, EKG interpretation). The students estimated their scores on each station. Self-assessment accuracy was determined for *each subject* through a correlation of each subject's self-ratings with the expert ratings' provided for that individual over the ten stations. Thus, each student's correlation with the expert over the ten stations reflects the extent to which that student was able to identify the stations at which he performed well and the stations at which he performed poorly *compared to his own performance at the other stations*.

The application of the intraindividual approach to the measurement of self-assessment across a set of skills within a single task was illustrated in a study by Gruppen et al. (1997). Students were evaluated on their performance of a standardized patient (SP) interview. The evaluation form contained seven items. The correlation coefficient between each student's self-assessments and the SP's assessments was calculated over the seven items. Again, each individual's correlation with the SP reflects her ability to identify the dimensions of performance on which she was relatively effective or ineffective relative to her performance on the other dimensions.

The major limitation of this new paradigm as described in the studies above is that it depends heavily on variation of scores. For example, suppose a student gives all aspects of his performance a 3 out of 5. Even if this represents an accurate self-assessment of performance (i.e. the student has correctly determined that all aspects of performance are moderate), the correlation between student and expert will be zero (because there is no variation in the multiple scores generated by the student). This may be a significant issue. One of the most common errors in global assessment is the halo effect, which describes the tendency of a rater to give similar evaluations to separate aspects of an individual's performance (Gray, 1996). Therefore, if this approach is employed in the study of self-assessment, it is important to ensure that subjects and experts are using the full range of the scale provided.

In an effort to circumvent this potential problem, Regehr and colleagues (1996) proposed a modified Q-sort method, which they termed the 'relative ranking model'. The Q-sort method has been used for the study of psychological characteristics (Bem and Funder, 1974; Block, 1978; Mowrer, 1953). Regehr and colleagues (1996) applied this technique to the evaluation of communication skills in a psychiatric interview. The evaluation form listed ten skills and a horizontal row of ten boxes. Above box 1 appeared the descriptor 'needs most work', above box 10 'needs least work', and boxes 5 and 6 were paired with the descriptor 'mid-level

skills'. Following a standardized patient encounter, students were asked to identify skills that belonged at the extremes and center of the scale, and then place the remaining skills using one item per box. The expert rated the performance using the same form. A correlation was calculated for each individual student's rankings of skills with each of the experts' rankings.

The authors concluded that the relative ranking model was not only useful as an alternative measure of self-assessment, but also ideal for the provision of feedback and could be used for educational purposes. However, as an assessment tool, this method has limitations. It is still important to study self-assessment as it pertains to the recognition of overall competence or incompetence, as this is key to safe practice. The Q-sort technology, however, is entirely intraindividual and therefore makes no effort to assess overall ability against some relevant criterion. In addition, this model may generate a somewhat 'artificial' hierarchy of strengths and weaknesses.

Summary

The importance of self-assessment in education stands undisputed. The health professions and higher education literature testify to this fact, as reflected in a sustained interest in this domain for over 30 years. Despite the theoretical value of self-assessment, the traditional measures employed in this literature could lead to the conclusion that self-assessment ability is poor. However, problems inherent in the traditional approaches for measuring self-assessment call into question this verdict on self-assessment.

In this paper, we have proposed several ways to address the methodological issues that arise within the traditional paradigms for studying self-assessment. For instance, the use of multiple expert raters can assess the reliability (or lack thereof) of the gold standard, the problem of differential use of the scale among students may be addressed by providing methods to 'calibrate' their self-evaluations, and the pitfalls as a cluster can be controlled by comparing self-assessment to peer assessment in the same individuals. In addition, we have offered one alternative framework for research in this domain that conceptualizes self-assessment as an intraindividual (as opposed to interindividual) process. This new framework faces its own limitations, but it exemplifies the potential that exists to take the study of self-assessment in new directions. We do not claim to have solved all the problems inherent in the self-assessment literature. However, we do strongly suggest that the current status quo is insufficient. Studies that make use of the traditional designs to study self-assessment without accounting for the potential methodological flaws inherent in these approaches will not be able to contribute meaningfully to the self-assessment literature in the future.

References

Abrams, R.G. & Kelley, M.L. (1974). Student self-evaluation in a pediatric-operative technique course. *Journal of Dental Education* **38**: 385–391.

Antonelli, M.A.S. (1997). Accuracy of second-year medical students' self-assessment of clinical skills. *Academic Medicine* **72**(10 Supplement 1): S63–S65.

Arnold, L., Willoughby, T.L. & Calkins, E.V. (1985). Self-evaluation in undergraduate medical education: A longitudinal perspective. *Journal of Medical Education* **60**: 21–28.

Bem, D.J. & Funder, D.C. (1974). Predicting more of the people more of the time: assessing the personality of situations. *Psychological Review* **8**: 506–520.

Bergee, M.J. (1997). Relationships among faculty, peer, and self-evaluations of applied performances. *Journal of Research in Music Education* **45**: 601–612.

Bishop, J.B. (1971). Another look at counselor, client, and supervisor ratings of counselor effectiveness. *Counselor Education and Supervision* **10**: 319–323.

Block, J. (1978). *The Q Sort Method in Personality Assessment and Psychiatric Research*. Palo Alto, CA: Consulting Psychologists Press.

Boud, D.J., Churches, A.E. & Smith, E.M. (1986). Student self assessment in an engineering design course: An evaluation. *International Journal of Applied Engineering Education* **2**(2): 83–90.

Boud, D.J. & Tyree, A.L. (1979). Self and peer assessment in professional education: A preliminary study in law. *Journal of the Society of Public Teachers of Law* **15**(1): 65–74.

Burke, R.J. (1969). Some preliminary data on the use of self-evaluations and peer-ratings in assigning university course grades. *Journal of Educational Research* **62**(10): 444–4448.

Calhoun, J.G., Ten Haken, J.D. & Wooliscroft, J.O. (1990). Medical students' development of self and peer assessment skills: a longitudinal study. *Teaching and Learning in Medicine* **2**: 25–29.

Calhoun, J.G., Wooliscroft, J.O., Ten Haken J.D., Wolf F.M. & Davis, W.K. (1988). Evaluating medical student clinical skill performance: Relationships among self, peer, and expert ratings. *Evaluation & the Health Professions* **11**: 201–212.

Chiu, L.H. (1975). Influence of student teaching on perceived teaching competence. *Perceptual and Motor Skills* **40**: 872–874.

Cochran, S.B. & Spears, M.C. (1980). Student self-assessment and instructors' ratings: A comparison. *Journal of the American Dietetic Association* **76**: 253–257.

Coutts, L. & Rogers, J. (1999). Predictors of student self-assessment accuracy during a clinical performance exam: comparisons between over-estimators and under-estimators of SP-evaluated performance. *Academic Medicine* **74**(10 Supplement): S128–S130.

Daniel, S.J., Scruggs, R.R. & Grady, J.J. (1990). Accuracy of student self-evaluations of dental sealants. *Journal of Dental Hygiene* **64**: 339–342.

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**: 411–418.

D'Augelli, A.R. (1973). The assessment of interpersonal skills: A comparison of observer, peer and self ratings. *Journal of Community Psychology* **1**: 177–179.

Davis, J.K. & Rand, D.C. (1980). Self-grading versus instructor grading. *Journal of Educational Research* **73**(4): 207–211.

Doleys, E.J. & Renzaglia, G.A. (1963). Accuracy of student prediction of college grades. *Personnel and Guidance Journal* **41**(6): 528–530.

Everett, M.S. (1983). Influence of trait anxiety on self-grading. *Educational Directions* **8**(1): 4–9.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment and Evaluation in Higher Education* **11**(2): 146–166.

Falchikov, N. & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research* **59**: 395–430.

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research* **70**: 287–322.

Farnill, D., Hayes, S.C. & Todisco, J. (1997) Interviewing skills: Self-evaluation by medical students. *Medical Education* **31**: 122–127.

Filene, P.O. (1969). Self-grading: An experiment in learning. *Journal of Higher Education* **40**: 451–458.

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**: 737–741.

Forehand, L.S., Vann, W.F. & Shugars, D.A. (1982). Student self-evaluation in preclinical restorative dentistry. *Journal of Dental Education* **46**(4): 221–226.

Fuqua, D.R., Johnson, A.W., Newman, J.L., Anderson, M.W. & Gade, E.M. (1984). Variability across sources of performance ratings. *Journal of Counselling Psychology* **31**(2): 249–252.

Gaier, E.L. (1961). Student self estimates of final course grades. *Journal of Genetic Psychology* **98**: 63–67.

Geissler, P.R. (1973). Student self-assessment in dental technology. *Journal of Dental Education* **37**: 19–21.

Gordon, M.J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine* **66**: 762–769.

Gordon, M.J. (1992). Self-assessment programs and their implications for health professions training. *Academic Medicine* **67**: 672–679.

Gray, J.D. (1996). Global rating scales in residency education. *Academic Medicine* **71**(1 Supplement): S55–S63.

Gray, T.G.F. (1987). An exercise in improving the potential of exams for learning. *European Journal of Engineering Education* **12**(4): 311–323.

Greenfield, D.G. (1978). Evaluation of music therapy practicum competencies: Comparisons of self and instructor ratings of videotapes. *Journal of Music Therapy* **15**(1): 15–20.

Gruppen, L.D., White, C., Fitzgerald, J.T., Grum, C.M. & Wooliscroft, J.O. (2000). Medical students' self-assessments and their allocations of learning time. *Academic Medicine* **75**: 374–379.

Gruppen, L.D., Garcia, J., Grum, C.M., Fitzgerald, J.T., White, C.A., Dicken, L. et al. (1997). Medical students' self-assessment accuracy in communication skills. *Academic Medicine* **72**(10 Supplement 1): S57–S59.

Gruppen, L.D., Baliga, S., Fitzgerald, J.T. et al. (1998). Do personal characteristics influence self-assessment accuracy? In, *Proceedings of the 8th International Ottawa Conference on Medical Education*, pp. 304–309. Philadelphia, PA.

Harrington, J.P., Murnaghan, J.J. & Regehr, G. (1997). Applying a relative ranking model to the self-assessment of extended performances. *Advances in Health Sciences Education* **2**: 17–25.

Hay, J.A. (1995). Investigating the development of self-evaluation skills in problem-based tutorial course. *Academic Medicine* **70**: 733–735.

Henbest, R.J. & Fehrsen, G.S. (1985). Preliminary study at the medical university of Southern Africa on student self-assessment as a means of evaluation. *Journal of Medical Education* **60**: 66–68.

Herbert, W.N.P., McGaghie, W.C., Droegemueller, W., Riddle, M.H. & Maxwell, K.L. (1990). Student evaluation in obstetrics and gynecology: self versus departmental assessment. *Obstetrics & Gynecology* **76**: 458–461.

Hodges, B., Regehr, G. & Martin D. (2001) Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. *Academic Medicine*; **76**(10 Supplement): S87–89.

Irvine, J.J. (1983). The accuracy of pre-service teachers' assessments of their classroom behaviors. *Journal of Research and Development in Education* **17**(1): 25–31.

Israelite, L. (1983). Adult student self-evaluation. *Performance and Instruction Journal* **22**(5): 15–16.

Johnson, D. & Cujec, B. (1998). Comparison of self, nurse, and physician assessment of residents rotating through an intensive care unit. *Critical Care Medicine* **26**: 1811–1816.

Kaiser, S. & Bauer, J.J. (1995). Checklist self-evaluation in a standardized patient exercise. *American Journal of Surgery* **169**: 418–420.

Keefer, K.E. (1971). Characteristics of students who make accurate and inaccurate self-predictions of college achievement. *Journal of Educational Research* **64**(9): 401–404.

Kegel-Flom, P. (1975). Predicting supervisor, peer, and self ratings of intern performance. *Journal of Medical Educcation* **50**: 812–815.

Keynan, A., Friedman, M. & Benbassat, J. (1987). Reliability of global rating scales in the assessment of clinical competence of medical students. *Medical Education* **21**: 477–481.

Kolm, P. & Verhulst, S.J. (1987). Comparing self and supervisor evaluations – a different view. *Evaluation & the Health Professions* **10**: 80–89.

LeBlanc, R. & Painchaud, G. (1985). Self assessment as a second language placement instrument. *TESOL Quarterly* **19**(4): 673–687.

Leichner, P. & Kalin, R. (1980). Results of the first Canadian psychiatric knowledge self assessment for residents. *Canadian Journal of Psychiatry* **25**(4): 281–289.

Linn, B.S., Arostegui, M. & Zeppa, R. (1975). Performance rating scale for peer and self assessment. *British Journal of Medical Education* **9**: 98–101.

Linn, B.S., Arostegui, M. & Zeppa, R. (1976). Peer and self assessment in undergraduate surgery. *Journal of Surgical Research* **21**: 453–456.

MacFadyen, J. & Turnbull, J. (1985). Effect of curriculum on student self-assessment of communication skills. In *Proceedings of the 6th Ottawa Conference of Medical Education and Assessment*, pp. 153–154, Toronto, ON.

Mast, T.A. & Bethart, H. (1978). Evaluation of clinical dental procedures by senior dental students. *Journal of Dental Education* **42**(4): 196–197.

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**: 1201–1206.

McGeever, P.J. (1978). Student self-grading in the introductory American politics course. *Teaching Political Science* **5**(3): 319–330.

Mihal, W.L. & Graumenz, J.L. (1984). An assessment of the accuracy of self-assessment of career decision making. *Journal of Vocational Behavior* **25**: 245–253.

Moore, D.E. & Cordes, D.L. (1992). Needs assessment. In A.B. Rosof & W.C. Felch (eds.), *Continuing Medical Education: A Primer*, 2nd ed., pp. 42–51. Westport, CT: Praegar.

Morton, J.B. & MacBeth, W.A. (1977). Correlations between staff, peer and self assessments of fourth-year students in surgery. *Medical Education* **11**: 167–170.

Mowrer, O.H. (1953). 'Q-technique' – description, history, and critique. In O.H. Mowrer (ed.), *Psychotherapy Theory and Research*, pp. 316–375. New York: Ronald.

Mueller, R.H. (1970). Is self-grading the answer? *Journal of Higher Education* **41**(3): 221–224.

Murstein, B.I. (1965). The relationship of grade expectations and grades believed to be deserved to actual grades received. *Journal of Experimental Education* **33**(4): 357–362.

O'Neill, G.P. (1985). Self, teacher, and faculty assessments of student teaching performance: A second scenario. *Alberta Journal of Educational Research* **31**(2): 88–98.

Palmer, P.B., Henry, J.N. & Rohe, D.A. (1985). Effect of videotape replay on the quality and accuracy of student self-evaluation. *Physical Therapy* **65**: 497–501.

Pease, D. (1975). Comparing faculty and school supervisor ratings for education students. *College Student Journal* **9**(1): 91–94.

Pitishkin-Potanich, V. (1983). On evaluating students' knowledge. *Higher Education in Europe* **8**(2): 18–22.

Plorde, D.S., Wollitzer, A.O. & Blossom, H. J. (1985). A self-assessment questionnaire for patient educators. *Journal of Medical Education* **60**: 728–730.

Pohlmann, J.L. & Beggs, D.L. (1974). A study of the validity of self-reported measures of academic growth. *Journal of Educational Measurement* **12**: 115–118.

Regehr, G., Hodges, B., Tiberius, R. & Lofchy, J. (1996). Measuring self-assessment skills: An innovative relative ranking model. *Academic Medicine* **71**(10 Supplement): S52–S4.

Rezler, A.G. (1989). Self-assessment in problem-based groups. *Medical Teacher* **11**: 151–156.
Risucci, D.A., Tortolani, A.J. & Ward, R.J. (1989). Ratings of surgical residents by self, supervisors, and peers. *Surgery, Gynecology & Obstetrics* **169**: 519–526.
Sclabassi, S.E. & Woelfel, S.K. (1984). Development of self-assessment skills in medical students. *Medical Education* **84**: 226–231.
Spencer, J.A. & Jordan, R.K. (1999). Learner centred approaches in medical education. *BMJ*. **318**: 1280–1283.
Stanton, H.E. (1978). Self-grading as an assessment method. *Improving College and University Teaching* **26**: 236–238.
Stover, R.V. (1976). The impact of self-grading on performance and evaluation in a constitutional law course. *Teaching Political Science* **3**(3): 303–310.
Stuart, M.R., Goldstein, H.S. & Snope, F.C. (1980). Self-evaluation by residents in family medicine. *Journal of Family Practrice* **10**: 639–642.
Turnbull, J., Gray, J. & MacFadyen, J. (1998). Improving in-training evaluation programs. *JGIM* **13**: 317–323.
Wheeler, A.E. & Knoop, H.R. (1981). Self, teacher and faculty assessments of student teaching performance. *Journal of Educational Research* **75**(3): 171–181.
Wooliscroft, 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**: 285–294.
Zonia, S.C. & Stommel, M. (2000). Interns' self-evaluations compared with their faculty's evaluations. *Academic Medicine* **75**: 742.