Reflections

Assessing Tutorial-Based Assessment

KEVIN W. EVA
Department of Clinical Epidemiology and Biostatistics, McMaster University, 1200 Main St. W.,
Hamilton, Ontario, Canada L8N 3Z5

Abstract. Since the development of problem-based learning curricula, medical educators have viewed tutorials as an ideal context within which medical student competence can be assessed. Advantages of tutorial-based assessment include, (a) evaluation based on prolonged and intense interactions between students, peers, and tutors, (b) the opportunity to assess domains of competence that are not readily assessed by more traditional examinations (e.g., communication skills, scientific curiosity, and respect for peers), and (c) avoidance of the negative impact of more formal summative evaluations. However, in addition to psychometric weaknesses inherent in this form of evaluation, the interpersonal relationships that serve as one of the primary strengths of the tutorial have also proven to be one of the main defects in tutorial-based assessment. The current paper will review the evidence relevant to this assessment paradox, paying particular attention to the use of self-assessment.

Key words: competence, evaluation, peer-assessment, self-assessment, tutor-assessment, tutorial-based assessment

Introduction

As medical knowledge becomes more rapidly generated and disseminated, an increasing emphasis has been placed on the need for physicians to develop the skills necessary for life-long learning. These skills include the ability to (a) recognize where one’s own knowledge is inadequate, (b) gather information relevant to the recognized deficiency, (c) critically appraise newly acquired information, and (d) implement new understanding in a manner that improves practice. The emphasis on the development of these skills was, in part, the impetus for the creation of curricula that utilize problem-based learning (PBL) (Barrows, 1983). While any attempt to further promote these skills is laudable, evaluating student progress in these areas has proven difficult. To maintain an educational system that is student-centered, assessment protocols within PBL curricula have often emphasized self- and peer-evaluation, usually within the context of small group tutorials. The current paper will review the evidence relevant to tutorial-based evaluation by examining the use of tutor-, peer-, and self-assessment. Beforehand, however, a closer look at the need for this type of assessment will be undertaken.
The Need for Tutorial-Based Assessment

The evaluation strategy implemented by medical schools has long been known to influence students’ study habits (Newble and Jaeger, 1983). As a result, care must be taken to ensure a match between curricular ideology and assessment procedures. As Norman (1991) has noted, generating detailed behavioural objectives as a precursor to an assessment exercise is the antithesis of problem-based, self-directed learning. Similarly, evaluation systems that encourage cramming for a formal examination do not fulfill the primary objective of PBL – promoting learning within the context in which the material will need to be retrieved in practice (Norman, 1988). Since PBL units are designed to emphasize principles of analysis and the resolution of clinical problems (Barrows and Tamblyn, 1976) the emphasis of the assessment tools should focus, at least in part, on the evaluation of the skills, processes, and attitudes of each student within the context of clinical issues. Intuitively, the most appropriate setting for implementing these evaluation ideals is the tutorial itself. Herein lies the paradox.

On one hand, tutorial-based assessment is intended to allow for evaluation based on actual performance to take place within the learning environment (Hay, 1995). Students should be able to assess themselves, as it is believed that students who maintain a more active role in the learning process (as is stressed in PBL tutorials) should be better able to judge their own performance. Furthermore, the dynamic nature of these groups and the prolonged interactions are expected to allow tutors and peers to develop a better sense of the strengths and weaknesses of the members of the group, thereby allowing them to provide accurate and constructive feedback. As a result of these apparent advantages McMaster University’s medical programme initially relied solely on tutorial-based assessment (Norman, 1991).

However, there is a real concern that conducting evaluation in the tutorial setting may negatively impact on the process itself. So, paradoxically, while the tutorial is the best place to observe these behaviours, the act of observation and evaluation may inhibit their demonstration. Thus, the dual roles of mentor and judge are viewed as incompatible by many tutors (Blake et al., 1995). The University of Maastricht has avoided the use of tutor-based assessment for this very reason (Nendaz and Tekian, 1999).

Similarly, the use of self- and peer-based assessment has been argued against. The intimate nature of the relationships that has often been viewed as positive for its ability to provide peers with an accurate understanding of their colleagues’ competence also makes students hesitant to rate their peers for the purposes of evaluation (Van Rosendaal and Jennett, 1992). Students are, not surprisingly, even less likely to rate themselves as incompetent. As an anonymous educator once noted “even the church never asked parishioners to make their confession in public” (Blake et al., 1995, p.899).

That being said, the tutorial remains a source of an abundance of information even if comfort in generating a formal assessment based on that information has
proven more difficult to achieve than first expected. While multiple examination formats have been shown to effectively measure knowledge, the tutorial remains the most frequent opportunity to assess other abilities such as the development of clinical reasoning, self-directed learning, communication, internal motivation, critical thinking, and the ability to work effectively in a team. As a result, tutorial-based assessment remains an important component of evaluation in the PBL setting and how these skills might best be assessed within the tutorial remains an important question.

Before reviewing the evidence that has been collected regarding the use of self-, peer-, and tutor-assessments, a few notes should be made. First, this introduction has focused primarily on tutorial-based assessment within a PBL curriculum, simply because much of the research that will be discussed arose from work done in a PBL context. This does not imply that the lessons learned do not apply to tutorials within more traditional environments. The constructs of critical thinking, self-direction, etc. are important for judging the competence of all physicians even though they have been emphasized primarily by educators working within PBL. Second, while knowledge can reliably be assessed using evaluation methods that are not tutorial-based, some of the research to be presented has focused on the evaluation of knowledge. However, the primary focus of this review will not be the assessment of any one skill in particular. Rather, the questions of interest are more generic. Can students self-assess their abilities? Are tutors or peers better at assessing the competence of their students/colleagues? Furthermore, evidence will be gathered from studies performed in multiple settings, only one of which will be tutorials. After all, as Hay (1995a) has argued, determining what to evaluate should be presaged by the question of whether to evaluate at all. If self-, peer-, or tutor-based assessments have fundamental flaws, then it makes no sense to examine the use of these assessments as they pertain to the evaluation of specific skills within tutorials. Finally, one additional question that could be raised is what impact does self-assessment have on subsequent performance? This is an important issue that is mostly beyond the scope of this review, but some evidence that is relevant to the topic will be raised indirectly in the section entitled self-assessment.

Tutor-based Assessment

Much of the research in self- and peer-based assessment uses tutor-based ratings as a gold standard. It is therefore important to consider the reliability of such ratings. So far it has been argued that one of the main strengths of tutorial-based assessment is that evaluation can be based on performance during prolonged interactions, thereby allowing the tutor to make a more accurate estimate of a student’s competence than might be provided during a more formal examination. This view appears to be simplistic. It has proven difficult for tutors to judge whether or not their students have acquired the desired competencies based on evidence obtained during tutorial sessions. One reason for this difficulty, posed by Graesser and
Person (1994), is that the questions asked during tutorial sessions are often open-ended, and therefore, do not provide tutors with the opportunity to assess individual levels of understanding or to draw comparisons across individuals. In an attempt to resolve this difficulty, a series of assessment instruments have been developed to assist tutors in evaluating their students within the tutorial setting.

Hebert and Bravo developed the first standardized instrument designed to assess the attitudes and skills acquired by medical students as a result of PBL tutorials (Hebert and Bravo, 1996). After administering a 44-item instrument to 273 students, factor analysis revealed four factors that could be evaluated independent of one another using their Tutotest: effectiveness, communication and leadership, scientific curiosity, and respect for peers. Averaging over seven evaluations, the reliability of Tutotest was 0.86 and Tutotest scores were moderately correlated with both the tutor’s global evaluation \( r = 0.64 \) and the students’ written examination results \( r = 0.39 \). Similarly, Valle et al. (1999) and Des Marchais and Vu (1996) have both developed instruments aimed at assessing student performance during PBL tutorials. Des Marchais and Vu performed their work at the Université de Sherbrooke and reported that both students and teachers were reluctant to use the tutor rating form. However, without reporting detailed evidence, they also claimed that in time, all participants discovered that the evaluation instrument does have a high likelihood of identifying academically weaker students, thereby lending validity to its use.

In summary, while it seems that tutor ratings based on tutorial performance can be useful in evaluating students’ skills, it is more difficult to generate reliable ratings than was first assumed. The apparent need to utilize standardized evaluation instruments demands additional involvement on the part of staff members to ensure that the resources are available for tutors, thereby lessening one advantage of tutorial-based assessment – the low cost and flexibility with which these ratings can be made. The cost will be higher still if content knowledge is a mandatory component of the tutor’s repertoire as was contended by Schmidt (1983). Nonetheless, the cost (both financial and temporal) of using standardized instruments is still likely to be lower than that incurred through the use of more elaborate examination protocols. Furthermore, the opportunity to measure factors that are not easily amenable to examination by more objective methods remains, and potential exists for the reliability of tutor-based assessments to be improved through the supplementary use of peer evaluation. Peers provide feedback to their colleagues on a regular basis and this information can potentially be useful to the tutor by improving upon the accuracy and comprehensiveness of the evaluations generated.

**Peer-based Assessment**

While tutor-based ratings are expected to be valid as a result of extended interaction between students and tutors, Van Rosendaal and Jennett have argued that the relationship between peers is even “more prolonged, intense, continuous, and
less subject to behavioral manipulations that may influence evaluation” (1994, p. 300). Furthermore, peer-ratings have the additional advantage of allowing an average score to be calculated across the opinions of many individuals, thereby providing for the generation of a more stable estimate of competence; global ratings of medical students, performance can provide a reliable assessment if a sufficient number of observers are used (Carline et al., 1992). Given the increased cost and difficulties associated with the addition of every tutor, peer ratings seem the optimal way to increase the number of observations. However, as alluded to earlier, the negative aspects of peer ratings have been cause for concern. Van Rosendaal and Jennett (1992) reported that residents in internal medicine felt very strongly that peer ratings are a major intrusion into the relationships among residents. This opinion did not arise out of concern for being rated incompetent by colleagues; participants in this study were more concerned with having to report on the weaknesses of others. In fact, 89% felt that it was more threatening to act as an evaluator than it was to be evaluated.

Although this resistance to peer-based assessment does exist, the empirical evidence available suggests that peer ratings are highly reliable and perhaps the most valid of the three types of assessment reported upon in this essay. Korman and Stubblefield (1971) examined the predictive ability of multiple independent variables, including grade point average, performance within a faculty, and peer ratings, for predicting future internship success. The correlations were, at best, in the low to moderate range of 0.3 to 0.4, thereby suggesting that a considerable amount of variance remains unexplained, but still, peer ratings were by far most predictive. Sullivan et al. (1999), upon reviewing the evidence relevant to peer ratings as a general construct, argued that these ratings are (a) good predictors of future performance, (b) internally consistent and reliable, and (c) provide information that is not received from other traditional methods of evaluation. Only Schwarz et al. (1994) have compared peer ratings in the tutorial to more objective outcome measures, and they too came to a similar conclusion. Relative to tutor and preceptor ratings, peer ratings were most highly correlated with scores on the NBME Shelf Examination, a test of knowledge gain.

Still, while the evidence in favour of using peer ratings seems strong, caution must be expressed regarding their use. The peer ratings collected for the studies outlined in the previous paragraph were not used for the purpose of summative evaluation of individual students. Hay has claimed that evaluations made by peers that have a direct impact on grades tend to be inflated and have a range that excludes anything below average (Hay, 1995). No evidence was documented, but given the resistance to evaluating peers reported by Van Rosendaal and Jennett (1992) this claim is plausible. However, if peers rank their colleagues in an order that is consistent with their actual competence, then as long as the “above average” portion of the scale has a reasonable range, the fact that all ratings are above average should matter little and high correlations should still be observed between peer ratings and future performance. Furthermore, a norm referenced assessment
system would still allow identification of those students who are at the low end of
the class to be identified so that remediation can be considered. More important,
the ability of students to assess their peers in a valid and reliable manner provides
evidence that students are aware of the skills and abilities that will promote success
later on in one’s career and are also capable of evaluating those abilities. The ques-
tion remains whether this ability to assess others is transferable to the assessment
of one’s self.

Self-assessment

Self-assessment is perhaps the most critical of all tutorial-based assessments,
because it must be considered not only as a potential format of evaluation, but also
as a skill in and of itself that forms an important component of one’s competence.
That is, as medical understanding is constantly updated, the ability to evaluate
one’s own deficiencies will strongly influence one’s ability to practice medicine
over a prolonged period of time. The development of these skills is of sufficient
importance that elaborate training protocols have been developed. For example,
Barrows and Tamblyn (1976) advanced a self-assessment unit during which they
had students work through a clinical problem, read up on the topic, observe an
expert performance, etc. While their hope was that these units would foster self-
assessment skills, no outcome measures were reported. Barrows and Tamblyn did
note that, when evaluating themselves, students rarely missed any of the concerns
raised by faculty evaluators, thereby suggesting that students were capable of evalu-
ating their own performance. This result was promising, but because so much
effort was required for the development of these units, they are no longer in use.
As a result, the tutorial has once again become the primary setting for this type of
evaluation.

A ubiquitous finding in the study of self-assessment is that self-ratings are
less reliable than those of medical supervisors (Gordon, 1991). Furthermore, self-
ratings tend to correlate less well with ratings assigned by the tutor than do
peer-tutor ratings (Sclabassi and Woelfel, 1984; Morton and MacBeth, 1977).
So, while students seem capable of evaluating the abilities of medical students
consistent with the expectations of the tutor, they appear to do so less well when
evaluating their own performance relative to when evaluating that of their peers.
More equivocal is whether students are more critical or more lenient toward
themselves relative to their tutors. For every study that shows students tend to over-
estimate their performance (Calhoun et al., 1988) there is another study reporting
that students are hesitant to assign themselves a high mark, even when it’s deserved
(Stuart et al., 1980). Furthermore, the use of technology, such as videotaping one’s
performance for later assessment has added little to the ability of medical students
to assess their own ability relative to the assessments provided by faculty members
(Palmer et al., 1985).
These results are both surprising and disconcerting given the basic assumptions that the individual learner is superior at identifying deficiencies in his or her understanding and that understanding develops by making note of areas in which the learner is deficient. One possible explanation is that students require time to develop their self-assessment skills and, therefore, improvement will only be observed once students gain more experience. That is, the relationship between the self and tutor-ratings might be expected to strengthen over time as students gain more experience with the feedback they receive from their tutors. Das et al.’s (1998) findings support this hypothesis; 80% of students and 70% of tutors believe that comparing tutor and self-evaluations is a beneficial exercise, thereby implying that a palpable skill can develop as a result of this interaction. However, this opinion-based data does not prove that the exercise of assessment actually influences self-evaluation.

Two studies have addressed this issue by examining the relationship between self and tutor assessments longitudinally. Arnold et al. (1985) examined self-evaluations in years 3, 4, 5, and 6 and found that, if anything, the relationship between student and tutor assessment became weaker with time. In comparing self-ratings to demographics, previous scholastic achievement, faculty ratings and previous self-assessments, they found that the most significant correlate of self-evaluations were, by far, previous self-ratings that students had generated. Among the concurrent correlates of self-ratings, the most consistent and significant were the docents’ ratings. However, as alluded to earlier, the strength of this relationship decreased with time and was non-significant by year six. Similarly, Rezler (1989) examined the correlations between self-ratings and tutor-ratings in multiple sections of units 1 and 6 of the medical program at the University of New Mexico. During unit 1, the two sets of scores correlated significantly in three out of four sections. However, later in time, during unit 6, the two sets of scores were unrelated on all four sections. Rezler hypothesized that students and faculty might rate students on different criteria. Specifically, she proposed that students rate themselves in terms of the tutorial group’s goals while faculty rate students in terms of factual knowledge. While intuitively plausible, this hypothesis, as stated, does not seem to account for the higher correlations between peer and tutor-ratings mentioned earlier, nor for the decrease in the strength of the relationship between self and tutor ratings across time. Still, this hypothesis will be considered in more detail shortly.

First, however, it should be noted that one longitudinal study did report an improvement in the relationship between self and tutor-ratings over time. Unfortunately, the interpretation of the results of this study is open to question. Calhoun et al. (1990) had students work through a clinical problem while their performance was videotaped. Afterward, performance was rated by the student actor, the student’s peers, and a faculty member using a behaviorally anchored checklist. It was found that students over-estimated their own performance relative to both the ratings assigned by peers and by tutors. Again, peer and tutor ratings were
similar. After three years, all three groups of participants were shown the same videotape of the student’s performance and asked to re-evaluate the student. After this three-year lag it was found that students now rated themselves in a way that was consistent with the ratings assigned by their peers and faculty. Calhoun et al. interpreted these results as evidence that students became better self-assessors, and speculated that this might result from higher expectations held by seniors. This conclusion is unwarranted given the procedure used. That is, it is unclear whether student self-assessments became more similar to tutor ratings as a result of an improved ability to self-assess or simply as a result of the lag between performance and assessment. When working through a problem the actor is privy to a lot of information that remains unavailable to external observers. For example, whether or not a particular diagnosis (or action) was considered is known only to the actor unless he or she comments on it aloud. After a three-year lag, students will no longer remember the mental operations that remained internal and, as a result, they would be expected to rate performance based solely on the actions apparent to an external observer even though the assessment is, by definition, self-assessment. So, evaluations should be more consistent with those performed by tutors regardless of whether or not the students have improved their ability to self-assess and we are left with the conclusion that self assessment ability does not improve with experience in medical school. That being said, it is still possible that training directed specifically at developing self-assessment abilities a la Barrows and Tamblyn (1976) can improve self-assessment. Such training does not appear to be naturally incorporated into medical curricula. Nonetheless, it is an idea worth pursuing as the number of opportunities available for specific and timely feedback from the clinical instructor likely decreases with time in medical training rather than increasing and hence, the need to accurately assess one’s own needs will increase.

Henbest and Fehrsen used evaluation guidelines, orientation sessions, discussions, etc. to ensure that both students and tutors were evaluating performance based on the same criteria (Henbest and Fehrsen, 1985). After doing so, they found strong agreement between the assessments provided by the two groups. The Spearman rank correlation was 0.74 and 80% of the students assigned themselves marks within 5% of the marks assigned by the faculty. Furthermore, the students’ identification of their own strengths and weaknesses appeared to be as thoughtful as those identified by faculty. Other training programs have reported similar results. For example, Cochran and Spears (1980) showed an improvement in the strength of the relationship between self- and tutor-based assessment upon identifying specific behaviours that students needed to improve upon and Abrams and Kelley (1974) found the same thing upon providing explicit and measurable criteria for judging dental preparations. Gordon (1992) reviewed eleven pioneering studies in which the authors attempted to evaluate the success of self-assessment programs and nine of the eleven reported cognitive and non-cognitive benefits to learners. As acknowledged by Gordon, these studies were imperfect, yet the consistency of their results...
held promise. In general, it seems that the more explicit the evaluation criteria, the more accurate are the self-assessments relative to assessment performed by tutors. “Student-instructor agreement is not improved by conventional supervised clinical training or by unchallenged self-reflection, but is improved under conditions of explicit criteria, intentional incentives, practice, and feedback in which data-driven self-assessments are formally reconciled with external tests and supervisors’ judgments” (Gordon, 1991, p. 768). However, whether or not this process is beneficial might be, at least in part, a philosophical decision.

As alluded to earlier, it has been proposed that individuals rate themselves, their peers, and their students based on different criteria. For example, perhaps students rate themselves based on past or ideal performance, whereas they rate peers based on current or actual performance. By examining experimental debriefing interviews, Stuart et al. (1980) did report evidence that residents in family practice tended to rate their “usual performances rather than the samples of behaviour that they were asked to evaluate”. However, Wooliscroft et al. (1993) have shown that self-assessment does not significantly correlate with prior academic performances. Blumenfeld et al. (1986) have reported a moderate to high correlation between self-assessments and the amount of effort expended, thereby suggesting that people have difficulty separating their actual performance from an ideal. These latter findings are consistent with Kegel-Flom (1975) and Risucci et al.’s (1989) argument that each observer (self, peer, and supervisor) contributes a unique perspective to the evaluation process, thereby resulting in poor correlations between them as performances are evaluated using different criteria. The factor analyses presented by these authors further support their argument as well as the hypothesis raised in explanation of Calhoun et al.’s findings that performance assessments might differ due to external viewers being forced to base their judgments on actions that cover a restricted range (i.e., what they can see) of student performance.

So, perhaps we should not strive to improve the low correlations between self and tutors as evaluation from multiple perspectives might be advantageous by providing a more holistic picture of one’s performance. Indirect evidence in favour of this arises from Hay’s (1995b) investigation of whether or not the process of self-assessment helps students develop a more general skill in self-evaluation. As the correlations between self and faculty assessment increased in a saw-tooth fashion, Hay was led to the conclusion that the improvement observed resulted from a “dance of negotiation” rather than an improvement in self-evaluation. It remains to be seen whether bringing self-assessment and the assessment of medical advisors in line with one another or maintaining a unique perspective better improves the ability to direct one’s own learning post medical school.

In summary, self-assessment ratings tend to correlate poorly with other measures of tutorial-based assessment unless explicit training in the evaluation criteria is provided. As the development of skills to allow self-direction is the goal of most self-assessment exercises, this is the outcome that should be held in highest regard. However, it is not yet clear which strategy (allowing students to assess from their
unique perspective versus teaching the criteria relevant to medical tutors) is most effective in nurturing the desired skill set.

Finally, a unique perspective regarding self-evaluation has been adopted by a pair of research teams. All of the studies discussed so far have examined self-evaluation as the ability to rate one’s own performance relative to that of one’s peers. However, in attempting to improve students’ ability to direct their own learning, it seems plausible that our goal should actually be to enable students to rank their own strengths and weaknesses relative to one another. That is, perhaps an intra-individual comparison is more important than an inter-individual comparison. Intuitively, such a focus could provide a rich formative assessment tool despite the fact that summative assessment is not possible with a ranking procedure. Gruppen et al. (1997) attempted to examine this possibility by asking students to rate themselves, using a 5-point scale, on 7 separate items relating to communication skills. There was no significant difference in the strength of the relationship between faculty and self-assessments between the intra-individual and the inter-individual approaches. However, as students were simply asked to rate themselves on multiple items, it is still possible that students rated themselves on each item relative to the perceived ability of their peers, thereby eliminating the change in focus that was desired.

In contrast, Regehr et al. (1996) utilized an intra-individual model by having students rank their own abilities relative to one another. Again, these researchers used the assessment of communication skills as their focus; ten skills were taken from a communication skills checklist and students were asked to order them from their own weakest ability to their own strongest ability. Upon doing so, Regehr et al. found a moderate correlation between student rankings and tutor rankings (0.58 when corrected for imperfect reliability of tutor rankings), thereby indicating that students might be better at evaluating their own skills relative to one another than they are at evaluating their own ability relative to that of their peers.

While these results yielded promise, it remains to be seen whether or not the relative-ranking procedure will provide a useful self-assessment tool for students in tutorials. More recent work suggests that the less constrained the context within which evaluation takes place, the less reliable the assessments tend to be, even when the relative ranking model is used. In reviewing the literature on self-assessment, Harrington et al. (1997) separated research that was based on long-term performance (e.g., post-clinical rotation) and that which was based on performance on a single task (e.g., an OSCE). In general, self-assessment is better correlated with other accepted measures of performance for short-term evaluations relative to long-term evaluations.

This dissociation has also been observed within the domain of relative-ranking exercises. While Regehr et al. (1996) observed a moderate correlation between self-assessments and observer assessments when psychiatric residents were asked to rank-order their performance on a single OSCE station, Harrington, Murnaghan and Regehr’s (1997) study of performance evaluation at the end of an ortho-
pedic rotation did not yield such promising results. They reported a corrected correlation between self-assessments and faculty ratings of 0.38 despite use of a relative-ranking model. Reiter et al. (in prep.) have also attempted to utilize a relative ranking model, this time to assist students in assessing their performance in the specific context of PBL tutorials. They observed no relationship between the rankings assigned by tutors, peers, or self-raters. Even more problematic is that the rankings assigned by individuals were not reliable across a two-week delay. Taken together, these studies cast doubt on whether the relative ranking model of assessment can solve the problem of unreliable, and hence invalid, self-assessments within a broadly defined context such as tutorials.

Summary and Conclusions

The need for tutorial-based assessment was discussed as it pertains to the maintenance of the philosophy of PBL and the need to nurture the self-directed learning capacities of physicians. In addition, it has been noted that the prolonged interactions between students, their tutorial leaders, and their peers yields a resource that seems ripe with information that can potentially help guide competency decisions. There remains two additional, perhaps more pragmatic, reasons to continue to develop assessment techniques within the tutorial setting. First, tutorial-based assessment is not only potentially useful in helping students identify areas in which they need to further their learning, but it also has the potential to provide feedback to the program itself regarding which areas of the curriculum need to be more fully developed. Similarly, tutors can use these types of assessments to guide their teaching practices by allowing them to tailor their lessons toward areas in which students require more guidance.

Second, in many higher education programs, medicine included, faculty and students deplore the emphasis on grades (Bender, 1969). Yet for legal and ethical reasons, it is important for the medical school to graduate only those students who prove themselves to be competent physicians. Given that “many students become overly involved in grade getting, often to the detriment of more fundamental educational values” (Korman and Stubblefield, 1971, p. 670), determining how to evaluate student competence without emphasizing grade getting during formal examination periods becomes an important issue. Since the beginning of PBL, peer and self-assessment has been expected to solve this dilemma (Blake et al., 1995). In fact, there is some evidence that suggests that tutorial-based assessment can do just that. Des Marchais and Vu (1996) have reported a decrease in the amount of competition between students as a result of decreasing the number of summative examinations and giving more weight to other dimensions, particularly those that are evaluated by PBL tutors. This was true even though the Université de Sherbrooke was the only school, at the time, in which tutors performed a summative evaluation of students’ skills within a group (Hebert and Bravo, 1996) – an important point given the discontent that many educators and students have
expressed concerning the conflict of interest resulting from adopting the roles of both teacher and evaluator.

Despite these reasons for maintaining tutorial-based assessment, it is still true that the psychometric properties of many of these evaluation methods have proven deficient. Peer-based assessment appears to be the most reliable and valid of the three techniques discussed, but the resistance toward peer-based assessment also seems to be strongest. It is as a result of these deficiencies that the vast majority of schools have opted to use tutorial-based assessment for formative purposes only. As long as this formative assessment allows for the suggestion of remediation when it is warranted, I see no problem with using this strategy. However, it should be noted that some of the recent findings described in this paper appear to suggest that the psychometric shortcomings of these evaluation techniques can be overcome by training students and tutors to assess based on the same criteria and by providing carefully constructed evaluation forms (Hebert and Bravo, 1996). On the other hand, as expressed in the preceding section, improving the agreement between self and tutor ratings may or may not be beneficial. Kegel-Flom’s (1975) argument in favour of the benefits that can be derived from multiple view-points calls into question whether or not the validity of the three assessment techniques described should be measured against one another.

More fundamentally, it should be questioned whether or not tutorial-based assessment should be held to traditional psychometric standards at all. The answer is undoubtedly affirmative if self-assessment is intended to factor into a student’s grade. However, the act of self-assessment should be viewed as an opportunity to develop an additional skill similar to practicing laproscopy or learning to critically appraise research projects, rather than it necessarily being viewed as a way of evaluating a student’s knowledge.

It has been argued that even this conception of the purpose of self-assessment is destined to be inaccurate since one can only recognize a performance/knowledge deficiency if he or she is knowledgeable of the domain being considered. While I am sympathetic with this idea, I find it to be an insufficient argument. Anyone who practices in a field such as medicine will be aware that research and understanding is continually being advanced. As such, a key component of self-assessment might be the ability to recognize that a significant amount of time has passed since last reviewing the literature on current treatments or causes of a particular disease, to name two examples. That is, while a physician might perform competently for many years, routine checks of current practice should be commonplace and the process of assessing one’s own abilities in medical school might provide the impetus to develop such techniques.

That being said, the idea that this “skill” can be developed in such a way that it will have long-term implications remains open to question. More research is required to determine whether or not experience with formal self-assessment has any behavioural or cognitive affects despite the common finding that the assessments generated do not tend to correlate with other measures of ability. In addition,
future work might focus more closely on the identification of intermediate variables that might mediate self-ratings.

In summary, much of the work performed on tutorial-based assessment systems has resulted in reports of poor psychometric properties and resistance on the part of the people being asked to evaluate. It seems that the resistance can be overcome as evaluators gain experience with the procedure to be used. Furthermore, part of the reason for the reports of poor validity may result from the use of inappropriate comparisons as well as an under-appreciation of the variety of perspectives that can provide meaningful insight into a student’s level of competency. If self-assessment is to be treated as a ‘skill,’ the development of that skill needs to be studied in terms of its relation to future practice rather than in terms of the ability to match an external observer’s opinion of one’s ability or knowledge. In the meantime, however, it appears that peer-based assessments and tutor-based assessments do maintain strong enough reliability that they might conscientiously be used for formative evaluation purposes (especially when standardized evaluation instruments like Tuto-test (Hebert and Bravo, 1996) are used to supplement the global ratings assigned). Further work needs to be done, however, to determine the impact of these assessment exercises on future performance.

References


Bender, R.M. Attitude toward grading systems used in medical education. Journal of Medical Education 44: 1076–1081.


