Advances in Health Sciences Education (2007) 12:19–33
DOI 10.1007/s10459-005-2327-z

Using a Sampling Strategy to Address Psychometric Challenges in Tutorial-Based Assessments

KEVIN W. EVA[1,*], PATTY SOLOMON[2], ALAN J. NEVILLE[3], MICHAEL LADOUCEUR[4], KARYN KAUFMAN[5], ALLYN WALSH[6], and GEOFFREY R. NORMAN[1]

[1]*Department of Clinical Epidemiology and Biostatistics, Program for Educational Research and Development, MDCL 3510, McMaster University, Hamilton, ON, L8N 3Z5, Canada;* [2]*School of Rehabilitation Science, IAHS 437, McMaster University, Hamilton, ON, L8N 3Z5, Canada;* [3]*Department of Medicine, Michael G. DeGroote School of Medicine, MDCL 3101, McMaster University, Hamilton, ON, L8N 3Z5, Canada;* [4]*School of Nursing, HSC 2J26, McMaster University, Hamilton, ON, L8N 3Z5, Canada;* [5]*School of Midwifery, MDCL 3109, McMaster University, Hamilton, ON, L8N 3Z5, Canada;* [6]*Department of Family Medicine, Program for Faculty Development, MDCL 3510, McMaster University, Hamilton, ON, L8N 3Z5, Canada (*author for correspondence, E-mail: evakw@mcmaster.ca; Phone: +1-905-525-9140 x27241; Fax: +1-905-572-7099)*

Received 5 May 2005; accepted 22 August 2005

Abstract. *Introduction*: Tutorial-based assessment, despite providing a good match with the philosophy adopted by educational programmes that emphasize small group learning, remains one of the greatest challenges for educators working in this context. The current study was performed in an attempt to assess the psychometric characteristics of tutorial-based evaluation upon adopting a multiple sampling approach that requires minimal recording of observations. *Method*: After reviewing the literature, a simple 3-item evaluation form was created. The items were "Professional Behaviour," "Contribution to Group Process," and "Contribution to Group Content." Explicit definition of these items was provided on an evaluation form. Twenty five tutors in five different programmes were asked to use the form to evaluate their students ($N = 169$) after every tutorial over the course of an academic unit. Each item was rated using a 10-point scale. *Results*: Cronbach's alpha revealed an appropriate internal consistency in all five programmes. Test–retest reliability of any single rating was low, but the reliability of the average rating was at least 0.75 in all cases. The construct validity of the tool was supported by the observation of increasing ratings over the course of the academic unit and by the finding that more senior students received higher ratings than more junior students. *Conclusion*: Consistent with the context specificity phenomenon, the adoption of a "minimal observations often" approach to tutorial-based assessment appears to maintain better psychometric characteristics than do attempts to assess tutorial performance using more comprehensive measurement tools.

Key words: context specificity, small-group learning, tutorial-based assessment

Introduction

Since its inception, problem-based learning has created a challenge for educators charged with designing a protocol for student assessment (Blake et al., 1995). The difficulty arises from the desire to establish a cooperative learning environment in which students are guided by the principles of adult learning (e.g., being self-directed) rather than being examination driven. Early on, this challenge led to the complete avoidance of examinations; McMaster University's medical school relied entirely on tutorial-based assessment with tutors, peers, and the students themselves determining whether or not students had learned what needed to be learned (Cunnington, 2001). The prolonged and intimate interactions that small group, tutorial-based learning provides, be it problem-based or not, continues to lead tutors to believe they can validly discriminate the good students from the struggling. Decades worth of evidence, however, suggest that this belief might be erroneous. As a result, a number of other evaluation techniques have become prevalent within problem-based educational programmes (Nendaz and Tekian, 1999; Blake et al., 1996). Still, the inherent steering function present in all assessment strategies provides sufficient motivation to maintain tutorial-based assessment protocols in programs that expect a significant portion of learning to take place within tutorials. This paper reports our attempts to overcome these problems by assessing the psychometric characteristics of tutorial-based assessments through adoption of a multiple sampling approach in which minimal observations are recorded often.

A brief overview of the evidence pertaining to tutorial-based evaluation

The first class to complete McMaster University's medical program graduated in 1972. A year later it was clear that tutors were unable to predict who would succeed on the national licensing examination of the Medical Council of Canada (LMCC). Mueller et al. asked seven assessors to review the files of five graduates from the Class of 1973 who had failed the LMCC and eight randomly selected controls (Mueller et al., Unpublished report). Only 47% of students were categorized correctly. Fourteen years later, McAuley and Woodward repeated the study with nine faculty reviewers and 20 students (13 who failed the LMCC and the seven top ranked students) (McAuley and woodward, Unpublished report). 85% (11/13) of the students who failed were rated as having an average or above average knowledge base. Only one was predicted to do "poorly" on the LMCC. A decade later, Blake, Norman, and Mueller showed that tutorial reports could not even predict in course knowledge attainment (Blake et al., Unpublished report).

It would be easy to argue that these outcome measures are inappropriate. Perhaps the information that can be gleaned from tutorial-based interactions is meaningful and important, but unrelated to knowledge acquisition. In fact, when Neville surveyed tutors and students, asking them to rate the extent to which each of 40 items could be assessed in tutorial, four factors emerged: (1) tutorial dynamics, (2) awareness of gaps in tutorial process, (3) knowledge, and (4) strategies for learning (Neville, Unpublished report). The mean ratings assigned to each domain decreased in that order from a high of 6.1 out of 7 for student respondents (5.9 for tutors) for tutorial dynamics to lows of 4.24 and 4.55, respectively, for learning strategies. Didyk and Keane performed a similar study two years later, eliciting similar results (Didyk and Keane, 1997). When students and tutors were asked "which qualities does tutorial evaluation measure," only 47% of students and 63% of tutors responded positively to "knowledge." 2% and 6%, respectively, responded positively to "ability to pass the LMCC." In contrast, "communication skills" and "professional behaviours" were endorsed by 74–94% of respondents across both groups.

In the last 10 years, however, it has become apparent that the problem with tutorial-based assessments is more fundamental than simply expecting these evaluations to provide too broad a focus. Emerging evidence has called into question the extent to which tutorial-based evaluations are reliable, let alone valid. A burgeoning literature on self-assessment has suggested that there is little agreement between students and tutors regarding how well one has performed in tutorial (Eva, 2001), or indeed, how well one has performed on any task (Eva and Regehr, 2005). More disconcerting, however, is that Reiter et al. used a relative ranking model in a tutorial-based assessment context, and found that tutor judgments were inconsistent from one week to the next (Reiter et al., 2002). The test–retest correlation across 6 weeks of rankings appeared higher for tutors than for peer- or self-rankings, but it maximally reached only 0.22.

In response to findings such as those outlined above, numerous groups have attempted to create more systematic and standardized means by which to collect tutorial-based ratings of performance. Hebert and Bravo, for example, developed a 44-item questionnaire called Tutotest (Hebert and Bravo, 1996). Factor analyses revealed that the items clustered together into four groups: (1) effectiveness in the group, (2) communication and leadership skills, (3) scientific curiosity, and (4) respect for colleagues, the first factor accounting for 61% of the total variance. This study reported one of the better correlations found in the literature between tutorial-based ratings and knowledge scores ($r = 0.39$). Unfortunately, the test–retest reliability of the tool was in the low end of the moderate range (0.46). Ladouceur, et al. have similarly created a 31 item tutorial-based assessment tool (Ladouceur et al., 2004). The internal consistency was high, but test-retest reliability was not assessed.

Overcoming context specificity

This issue of poor correlation across assessments has been seen before (Eva, 2003). In fact, it is likely the most robust finding in health sciences education. Performance during a single clinical encounter is typically a poor predictor of performance on a second clinical encounter. The data reported by Reiter et al. suggest similarly, that performance in a single tutorial is a poor predictor of performance in a second tutorial (Reiter et al., 2002). That in and of itself is not a problem for those charged with designing assessment strategies; it simply mandates the collection of multiple observations because an individual's average performance across many problems can provide a reliable and valid estimate of ability even when context specificity rules (Swanson et al., 1995). The problem arises from the fact that most tutorial-based assessment strategies require tutors to complete an evaluation form at mid and end-unit, weeks of interaction often passing between recorded assessments. This protocol leaves tutors susceptible to recall bias in that their assessments will be most heavily guided by highly salient (i.e., memorable) events including early or recent tutorials. In other words, a tutor's memory may not adequately reflect each student's average performance over the course of a term. Memory biases such as these are known to lead to erroneous beliefs (Gilovich, 1991); we suspect they may be the cause of false inferences regarding one's ability to judge tutorial performance.

The creation of comprehensive tutorial-based assessment tools, while laudable in terms of the goal, may, therefore, constitute an infeasible strategy for overcoming the problems with tutorial-based assessments. Hebert and Bravo reported that five evaluations would be required for Tutotest to achieve an appropriate level of reliability. Busy clinical tutors, however, are unlikely to take the time to complete a 44-item questionnaire on every student in their tutorial five times over the course of an educational unit. As a result, we have adopted the opposite approach – to provide tutors with a minimal rating task, but ask them to complete the task after every tutorial rather than simply at mid-unit or end-unit. More extensive mid and end-unit evaluations are still completed, but are to be informed by the ratings/comments assigned after each tutorial. The remainder of this paper reports the development and testing of this novel "minimal observations often" approach to tutorial-based assessment.

Methods

Tool development

The authors of this paper constituted a Faculty-wide task force on tutorial-based assessment, representing multiple health sciences educational

programs. The large cumulative amount of experience with tutorial evaluation specifically (and educational principles/student assessment generally) maintained by the task force made the item generation phase inherently a content validity verification. Upon reviewing the literature outlined above, the task force deemed it appropriate to focus tutorial-evaluation efforts on three key domains:

(1) Professional Behaviours (i.e., attends tutorial, punctual, shows respect, provides/receives constructive feedback, demonstrates accountability),
(2) Contribution to Group Process (i.e., contributes to the development of objectives, completes negotiated tasks, encourages participation, identifies strengths and weaknesses, communicates effectively),
(3) Contribution to Group Content (i.e., clarifies points and enhances understanding, checks accuracy/validity of information, analyses/applies relevant theories/concepts/facts, generates/considers alternative perspectives, makes links with prior readings/experience/knowledge).

The phrase "contribution to group content" was adopted to acknowledge that assessing knowledge gain is particularly difficult in tutorial (and is better left to other forms of evaluation), but that students should still feel obligated to come to tutorial prepared to contribute information to the discussion. The general aim was to avoid forcing tutors to judge that which can not be judged; for example, having to determine, each time a student is quiet, whether that particular student feels uncomfortable with the particular content or simply does not want to dominate the conversation.

In addition to these three domains, it was decided that individual programs could add a fourth domain if it was felt that something was missing that constituted a particularly important piece of their tutorial-based curriculum, thereby allowing flexibility while still promoting adoption of a universal questionnaire. Physiotherapy, for example, opted to add "Evidence-based practice" as a focal domain.

Rather than creating a comprehensive list of questions, the form that was created consisted simply of 3 (or 4) 10-point rating scales, one corresponding to each of the domains listed above. The specific adjectives assigned to the 10 points varied as a function of program to enable the scale to correspond to the marking scheme accepted within each program and required for student records/transcripts. An example is provided in Appendix A. In addition to the quantitative rating scales, a comments page/box was included along with each set of scales. It was anticipated that completion of the form, which requires rating the performance of all students, would require approximately five minutes after each tutorial.

Setting and participants

Each of the programs represented in the task force (physiotherapy, nursing, midwifery, medicine, and health research methodology) recruited tutors and students to participate in a pilot study of this new rating protocol. Tutors were told (a) use of the form should not preclude delivery of qualitative verbal feedback, but rather, should constitute a record of the formative feedback delivered during the normal course of the educational unit, (b) that the domains they should consider were defined as above, but that the definitions should not be deemed comprehensive – rather, the definitions were provided simply as a guide to what should be considered, (c) they should assign a rating to each domain after every tutorial, and (d) they should feel free to share the form with students to provide them with a better sense of what is expected of them during tutorial. The number of participants and the characteristics of tutorial groups within each program are listed in Table I.

Analysis

Internal consistency and test–retest reliability were analyzed separately for each course using Generalizability Theory (Streiner and Norman, 2003). In all cases students were treated as the facet of differentiation. Test–retest reliabilities are reported for both a single observation and for the average of all observations provided to each group of students. Construct validity was assessed by comparing the mean scores (with 95% confidence intervals) received by students as a function of tutorial number (i.e., week of the course) and year in the programme (for physiotherapy, the only programme within which participants were recruited from multiple years of study). Our hypothesis in both cases was that, as students learn what is expected of them within tutorial, their performance should improve and the ratings tutors assign should increase as a result.

Results

The mean, standard deviation, and reliabilities associated with six independent replications, spread across five programs, are illustrated in Table II. The courses are ordered as a function of sample size, suggesting that the trustworthiness of each result descends as one reads down Table II.

Despite differences in sample size, some general trends are evident. First, as is the norm with tutorial-based evaluations (and, in fact, global performance assessments in general) the data are negatively skewed, most individuals receiving ratings at the end of the scale corresponding with a high level of performance. Mean scores were generally at least 8 out of 10 with a standard deviation of 1 point. Comparison of means across program would be inappropriate as there are too many contextual differences that could influence the relative standing of courses in our sample.

Table I. Characteristics of the sample/tutorial groups within each participating program

Program	Course	#Tutors	#Students	Students per group	Length of courses	Tutorials per course	Notes
Physiotherapy	Unit 2	8	55	6–7	12 weeks	12	– Two 2.5 hour tutorials per week. – Second unit of a six unit, 2 year professional masters curriculum. – Focuses on musculo-skeletal practice.
Medicine	Unit 2	8	47	5–6	12 weeks	24	– Two three-hour tutorials per week. – Second of four pre-clinical, problem-based, undergraduate curricular units. – Covers cardiovascular, respiratory, and renal systems.
Physiotherapy	Unit 5	5	31	6–7	12 weeks	12	– Two 2.5 hour tutorials per week – Fifth unit of a six unit, 2 year professional masters curriculum. – Focuses on community practice.
Health Research Methodology	727	2	16	8	12 weeks	12	– One three-hour tutorial per week. – Discussion-based, graduate-level course. – Introduction to scale development and reliability/validity testing.
Midwifery	Midwifery Care I	1	10	10	11 weeks	11	– One three-hour tutorial per week. – First of five clinical placements completed in second year of four year program.
Nursing	4Q04	1	10	10	13 weeks	13	– One three-hour tutorial per week, the form being completed seven times. – 4th year course in four year program focused on application of theoretical concepts to clinically encountered problems.

Table II. Means, standard deviations, and reliabilities of ratings assigned in tutorial as a function of course

Program	Course	Number of tutorials (n)	Professional behaviour		Group process		Group content		Internal Consistency (Alpha)	Test–Retest G(1)	Test–Retest G(n)*
			Mean	SD	Mean	SD	Mean	SD			
Physiotherapy	Unit 2	12	8.8	1.2	8.1	1.2	8.1	1.2	0.75	0.33	0.87
Medicine	Unit 2	24	9.0	0.6	8.8	0.8	8.8	0.8	0.83	0.30	0.92
Physiotherapy	Unit 5	12	9.4	0.8	8.6	1.0	8.6	1.1	0.80	0.18	0.75
Health Research Methodology	727	12	8.6	1.0	8.1	1.1	8.0	1.2	0.67	0.22	0.76
Midwifery	Midwifery Care I	11	7.9	1.3	7.7	1.3	7.7	1.6	0.94	0.65	0.95
Nursing	4Q04	7	6.6	2.0	6.5	1.9	6.6	2.0	0.94	0.52	0.88

*G(n) denotes the generalizability of the average of n tutorial evaluations, n being the number of tutorials within the course.

Second, this negative skew did not impact upon the ability of the form to discriminate between students. The internal consistency, a weak test of a scale's reliability (i.e., its ability to consistently discriminate between students), should ideally lie between 0.7 and 0.9, thus indicating a correlation between questions without excessive redundancy (Streiner and Norman, 2003). Alpha for the three samples of largest size fell precisely within this range and narrowly missed in the three samples of smallest size. More importantly, generalizability theory analyses were used to determine the extent to which students could be reliably discriminated from one another across tutorial. The test-retest reliability for a single observation (i.e., a single tutorial rating: G(1)) was quite poor, confirming the specific findings of Hebert and Bravo (1996) and the general phenomenon of context specificity (Eva, 2003). In contrast, the test–retest reliability of the average of all observations collected for each student does appear to provide a generalizable indication of each student's performance. G(n), 'n' being the number of observations collected per student, was at least 0.75 in all six samples, thus supporting our hypothesis that a multiple biopsy approach to tutorial-based assessment is both beneficial and required.

One program (physiotherapy) collected data from multiple units, thus allowing for a test of construct validity. Physiotherapy is a two year Master's level program, Unit 2 taking place in first year and Unit 5 taking place in second year. We hypothesized that, if students are developing the skills they are expected to develop within tutorial, then performance ratings using our scale should reflect this improvement, both within unit and across unit. Figure 1 illustrates the performance ratings for Unit 2 students as a function of week in tutorial and reveals a definite trend towards higher scores being assigned later in the unit. Figure 2 illustrates the performance ratings as a function of week for both Unit 2 and Unit 5 students, both lines representing an average taken across the four domains for which ratings were collected. The performance gains made in Unit 2 appear to have been maintained until Unit 5, Unit 5 students (i.e., those who have had greater experience in tutorial-based learning environments) having received higher performance ratings than did Unit 2 students early in the unit. Unit 5 students appear to have approached ceiling. Since the completion of this study, equivalent performance differences have been observed in the undergraduate MD program, Unit 4 students outperforming their more junior colleagues.

Discussion

When designing systems for student assessment, health professional training programs maintain responsibilities to the students (i.e., fairness and transparency), the curriculum (i.e., consistency of philosophy), and society (i.e.,

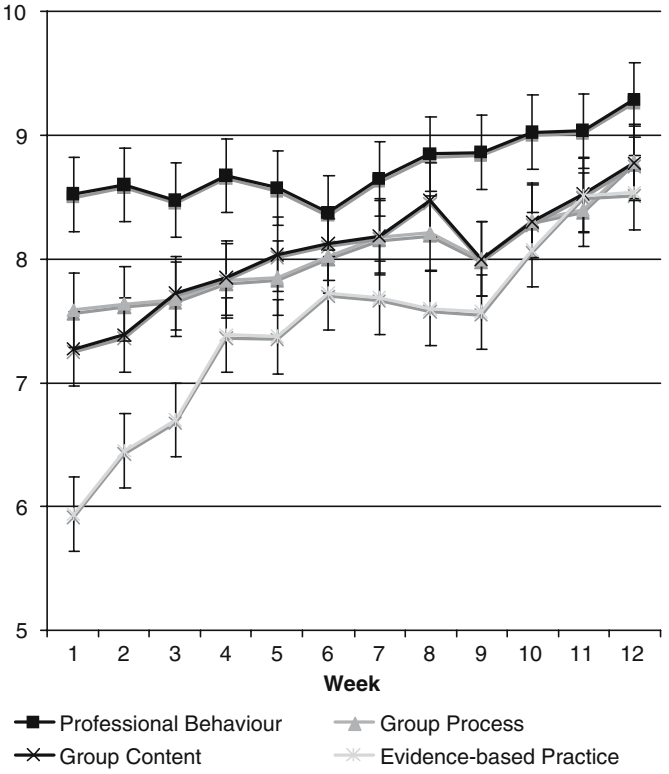

Figure 1. Mean performance ratings (and 95% confidence intervals) assigned to students in the physiotherapy program as a function of week within the course.

ensuring only competent students graduate to practice). Doing so requires finding a balance between reliability, validity, feasibility, and acceptability – the four "-ities" of good measurement. Tutorial-based assessments nicely fulfill the need to emphasize tutorial-based components of the curriculum and they provide a low-cost source of information regarding student performance, but numerous studies have suggested these forms of evaluation are psychometrically flawed. This blemish is not simply an academic issue. At McMaster, when the undergraduate MD program relied solely on tutorial-based evaluation, students were failing the national licensing exam and the program was unable to provide appropriate remedial assistance because the tutors were unable to predict which students were likely to have difficulty (Blake et al., 1995). The development and implementation of a more psychometrically sound assessment protocol that could provide students with reliable and valid feedback regarding their progress proved to be an important factor in reversing that trend (Blake et al., 1996). Still, tutorial-based evaluations fit well with the student-centred, cooperative learning

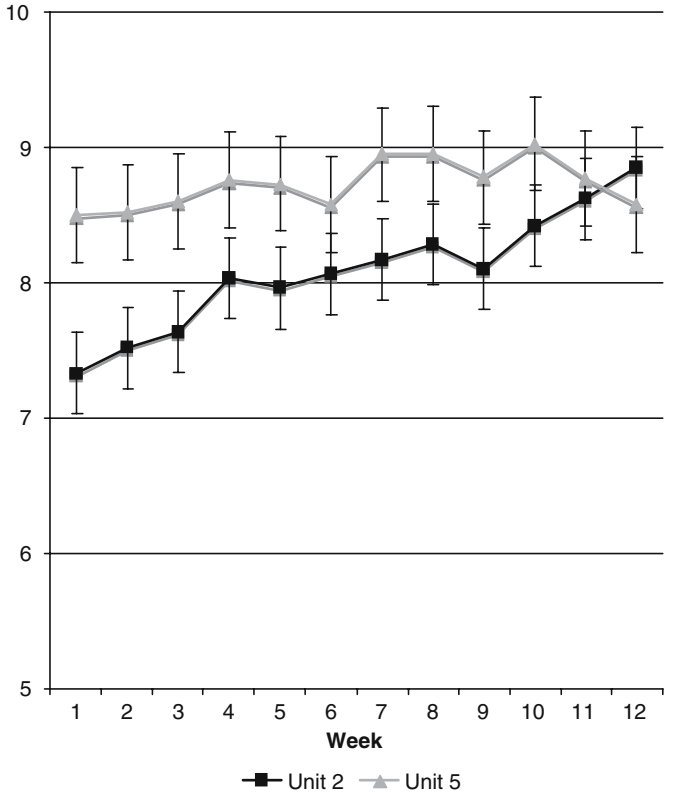

Figure 2. Mean performance ratings (and 95% confidence intervals) assigned to students in the physiotherapy program as a function of week and unit (averaged across domain.)

environment strived for at McMaster and, as Cunnington has noted, ''like any belief, the belief in the power of tutorial evaluation has been resilient and at times, even sacrosanct'' (Cunnington, 2001).

The data presented in this paper support the hypothesis that one way of bringing reality in line with belief is to limit the focus of tutorial-based evaluations to observable tendencies and to overcome context specificity by adopting a ''minimal observations often'' approach to measurement. Others have tried to scale-tweak by more systematically listing all of the skills and attitudes that might be evaluated in tutorial. Analogous to the conclusions of Schuwirth and van der Vleuten (Schuwirth and van der Vleuten, 2004), we view the stimulus format itself (in this case, frequency of sampling) as more important to the value of the tool than the response format (in this case, what questions are rated explicitly). Tutors simply can not be expected to assess dozens of student qualities for each of multiple students with any sort of regularity. As a result, the perceived advantage

provided by a full delineation of skills is overcome by the memory biases that arise from only being able to complete evaluation forms on one or two occasions. This is not to say that more careful listings of skills that could be considered within tutorial are useless. On the contrary, we suspect that dissemination of such listings could provide an educational opportunity for tutors regarding what variables should be considered when judging tutorial performance. Rather than expecting an explicit rating for each one, however, we would advocate simply rating the broader categories that encompass the specific items, so that ratings/comments can be assigned on multiple occasions.

The high internal consistency and low test–retest reliability found by Hebert and Bravo and replicated here in a different context supports this notion that "occasion" is a more critical source of error variance than is "item." If a measure's internal consistency is as high as the 0.98 reported by Hebert and Bravo, it suggests that not all items are required as the ratings contain sufficient redundancy as to allow equivalent information to be collected with less effort. We do not view the specific domains included in our minimized tutorial evaluation instrument (or the adjectives included with each scale) as magical or ideal – they simply fit well within the local context in McMaster University's Faculty of Health Sciences.[1] Other domains might be more appropriate in other institutions. The key to the success of our approach, rather, is ensuring that the domains are sufficiently limited in number to allow observations to accumulate with greater frequency than is the norm. In fact, if nothing else, we hope that the protocol we have developed will serve simply as a point of discussion among tutors, highlighting the need for careful and continuous record keeping on their part analogous to that of the patient record keeping expected of all health professionals.

Admittedly, the addition of a weekly record-keeping requirement was met with resistance from some tutors due to the perceived increase in workload the tool would create. However, many who used the form reported, anecdotally, that the small weekly increase in reporting time was more than offset by the reduced time and effort required to create formal mid-unit and end-unit evaluations. Others indicated an increase in comfort with their final evaluations and with their ability to counsel students on how to improve their performance over the course of the term. Others still were critical of the suggestion that their personal experience may not allow them to accurately track the performance of individuals without such systematic record keeping. To this we respond simply by drawing attention to the social psychology literature that shows lack of feedback can contribute to the creation and maintenance of erroneous beliefs (Gilovich, 1991). Only after systematically seeking data indicating how well our students performed on tasks outside of the tutorial environment were we startled to the

realization of just how poor our tutorial-based predictions could be. The data presented here suggest that systematically recording multiple samples of performance can lead to improvements in the psychometric properties of tutorial-based assessment.

Conclusion

In describing the evolution of evaluation at McMaster, Cunnington wrote "change must continue as we struggle to remain true to our roots while providing the best and most modern ... education possible" (Cunnington, 2001). Tutorial-based learning remains a core component of many health sciences curricula. To emphasize its importance and steer student effort, some form of tutorial-based evaluation is required. To do so, we recommend adopting a "minimal observations often" approach within which tutors are encouraged to keep systematic records, highlight observable behaviours, and provide explicit ratings to a core set of behavioural domains. These records should be informed by and used to inform the regular delivery of qualitative formative feedback.

Notes

[1] One caveat to this discussion is the use of a 10-point rating scale. Global ratings tend to cluster in the upper end of the scale. This clustering did not impede our ability to reliably discriminate between individuals because the upper end of 10-point scales provide a larger number of response options relative to shorter scales. This follows from a general principle of measurement that reliability increases with number of response options, counter to many people's intuitions. (Streiner and Norman, 2003).

Acknowledgements

The authors thank Christel Woodward, Linda O'Mara, Marilyn McIntyre, Mary-beth Ribble, and Wendy Edge for provision of some of the data included in this report. They also thank Susan Denburg for providing the impetus to pursue this project.

Appendix A

Sample response sheet, used by tutors in the Physiotherapy program (Domain definitions provided on cover sheet as described in methods section)

Student name	Domain	Rating Unsatisfactory/Satisfactory									
1._____.	1. Professional behaviour	1	2	3	4	5	6	7	8	9	10
	2. Contribution to group process	1	2	3	4	5	6	7	8	9	10
	3. Contribution to group content	1	2	3	4	5	6	7	8	9	10
	4. Evidence based practice	1	2	3	4	5	6	7	8	9	10
2._____.	1. Professional behaviour	1	2	3	4	5	6	7	8	9	10
	2. Contribution to group process	1	2	3	4	5	6	7	8	9	10
	3. Contribution to group content	1	2	3	4	5	6	7	8	9	10
	4. Evidence based practice	1	2	3	4	5	6	7	8	9	10
3._____.	1. Professional behaviour	1	2	3	4	5	6	7	8	9	10
	2. Contribution to group process	1	2	3	4	5	6	7	8	9	10
	3. Contribution to group content	1	2	3	4	5	6	7	8	9	10
	4. Evidence based practice	1	2	3	4	5	6	7	8	9	10
4._____.	1. Professional behaviour	1	2	3	4	5	6	7	8	9	10
	2. Contribution to group process	1	2	3	4	5	6	7	8	9	10
	3. Contribution to group content	1	2	3	4	5	6	7	8	9	10
	4. Evidence based practice	1	2	3	4	5	6	7	8	9	10
5._____.	1. Professional behaviour	1	2	3	4	5	6	7	8	9	10
	2. Contribution to group process	1	2	3	4	5	6	7	8	9	10
	3. Contribution to group content	1	2	3	4	5	6	7	8	9	10
	4. Evidence based practice	1	2	3	4	5	6	7	8	9	10
6._____.	1. Professional behaviour	1	2	3	4	5	6	7	8	9	10
	2. Contribution to group process	1	2	3	4	5	6	7	8	9	10
	3. Contribution to group content	1	2	3	4	5	6	7	8	9	10
	4. Evidence based practice	1	2	3	4	5	6	7	8	9	10
7._____.	1. Professional behaviour	1	2	3	4	5	6	7	8	9	10
	2. Contribution to group process	1	2	3	4	5	6	7	8	9	10
	3. Contribution to group content	1	2	3	4	5	6	7	8	9	10
	4. Evidence based practice	1	2	3	4	5	6	7	8	9	10

References

Blake, J.M., Norman, G.R. & Smith, E.K. (1995). Report card from McMaster: Student evaluation at a problem-based medical school. *Lancet* **345**: 899–902.

Blake, J.M., Norman, G.R., Keane, D.R., Mueller, C.B., Cunnington, J. & Didyk, N. (1996). Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine* **71**: 1002–1007.

Cunnington, J. (2001). Evolution of student evaluation in the McMaster MD Programme. *Pedagogue: Perspectives on Health Sciences Education* **10**: 13–18. http://www.fhs.mcmaster.ca/perd/download/Pedagogue_101.pdf.

Didyk, N. & Keane, D.R. (1997). Student and tutor opinion on five MD program evaluation tools. *Pedagogue: Perspectives on Health Sciences Education* **7**: 5–8.

Eva, K.W. (2001). Assessing tutorial-based assessment. *Advances in Health Science Education* **6**: 243–257.
Eva, K.W. (2003). On the generality of specificity. *Medical Education* **37**: 587–588.
Eva, K.W. (2005). Regehr G. Self-assessment in the Health Professions: A Reformulation and Research Agenda. *Academic Medicine* **80**(10suppl.): S46–S54.
Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press.
Hebert, R. & Bravo, G. (1996). Development and validation of an evaluation instrument for medical students in tutorials. *Academic Medicine* **71**: 488–494.
Ladouceur, M.G., Rideout, E.M., Black, M.E.A., Crooks, D.L., O'Mara, L.M. & Schmuck, M.L. (2004). Development of an instrument to assess individual student performance in small group tutorials. *Journal of Nursing Education* **43**: 447–455.
Nendaz, M.R. & Tekian, A. (1999). Assessment in problem-based learning medical schools: A literature review. *Teaching and Learning in Medicine* **11**: 232–243.
Reiter, H.I., Eva, K.W., Hatala, R.M. & Norman, G.R. (2002). Self and peer assessment in tutorials: Application of a relative ranking model. *Academic Medicine* **77**: 1134–1139.
Schuwirth L.W.T. & van der Vleuten C.P.M. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Medical Education* **38**: 974–979.
Streiner, D.L. & Norman, G.R. (2003). *Health measurement scales: A practical guide to their development and use,* 3rd edn. Oxford: Oxford University Press.
Swanson, D.B., Norman, G.R. & Linn, R.L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Research* **24**: 5–11,35.