Original Article

A Simple Instrument for the Assessment of Student Performance in Problem-based Learning Tutorials

Si-Mui Sim,[1] *BSc, PhD (UK)*, Nor Mohd Adnan Azila,[2] *BSc, PhD (UK)*, Lay-Hoong Lian,[2] *BSc, PhD (Mal)*,
Christina PL Tan,[3] *MBBS, MRCGP (UK)*, Nget-Hong Tan,[2] *BSc (Taiwan), PhD (USA)*

Abstract

Introduction: A process-oriented instrument was developed for the summative assessment of student performance during problem-based learning (PBL) tutorials. This study evaluated (1) the acceptability of the instrument by tutors and (2) the consistency of assessment scores by different raters. Materials and methods: A survey of the tutors who had used the instrument was conducted to determine whether the assessment instrument or form was user-friendly. The 4 competencies assessed, using a 5-point rating scale, were (1) participation and communication skills, (2) cooperation or team-building skills, (3) comprehension or reasoning skills and (4) knowledge or information-gathering skills. Tutors were given a set of criteria guidelines for scoring the students' performance in these 4 competencies. Tutors were not attached to a particular PBL group, but took turns to facilitate different groups on different case or problem discussions. Assessment scores for one cohort of undergraduate medical students in their respective PBL groups in Year I (2003/2004) and Year II (2004/2005) were analysed. The consistency of scores was analysed using intraclass correlation. Results: The majority of the tutors surveyed expressed no difficulty in using the instrument and agreed that it helped them assess the students fairly. Analysis of the scores obtained for the above cohort indicated that the different raters were relatively consistent in their assessment of student performance, despite a small number consistently showing either "strict" or "indiscriminate" rating practice. Conclusion: The instrument designed for the assessment of student performance in the PBL tutorial classroom setting is user-friendly and is reliable when used judiciously with the criteria guidelines provided.

Ann Acad Med Singapore 2006;35:634-41

Key words: C-IP skills, Consistency, Criteria guidelines, Intraclass correlations, Rating scores

Introduction

Assessment can be done in a variety of ways, for many purposes, and for different populations. It can occur at the classroom level, programme level, college level or even national level. It can take the form of *paper-and-pencil tests*, such as the multiple-choice question (MCQ) test, or *performance-based tests*, such as objective structured clinical examination (OSCE) or portfolio compilation. Since assessment is known to drive student learning,[1,2] the assessment method and the assessment instrument used can influence what and how students learn.[3,4] An appropriate assessment approach should serve to assist learning, measure individual achievement and provide valuable information on the implementation of a programme. While the assessment instrument selected must satisfy the intended educational outcomes of the programme, it should also be valid and reliable. For any assessment to be meaningful and representative of the student's true performance, the assessment scores should be reproducible. This reproducibility is usually not difficult to achieve with objective tests such as MCQ tests. However, such reproducibility becomes increasingly difficult to obtain with increasing variables (e.g., human subjectivity and time constraint) as found in written tests such as short answer questions (SAQ), modified essay questions (MEQ) and essays, and even more so with competency-based assessments,[2] such as OSCE and problem-based learning (PBL) tutorial assessments. In performance-based assessments, the measures taken to reduce variability of scores by the different raters include having checklists of

[1] Department of Pharmacology
[2] Department of Molecular Medicine
[3] Department of Primary Care Medicine
Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
Address for Reprints: Dr Debra Si Mui Sim, Department of Pharmacology, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia.
Email: debrasim@um.edu.my

items to be scored on, setting criteria for scoring each of the items and training raters for the task.

The New Integrated Curriculum (NIC) of the MBBS course at the University of Malaya, which was introduced in 1998/1999, is an integrated organ system-based curriculum. The curriculum is mainly lecture-based but incorporates elements of PBL, including small group PBL tutorials in the pre-clinical[5-7] as well as in the early clinical years. Some of the aims of including elements of PBL in the NIC are to induce students to improve their skills in communication, leadership and team-building, critical thinking or reasoning, and information-management.[8,9]

As a first step towards the development of a skills assessment system for a semi-integrated hybrid programme, a process-oriented instrument was developed for the summative assessment of student performance during the PBL tutorials. Assessment of processes and attitudes during tutorial sessions is considered to embody PBL principles and is the central focus of student assessment.[10,11]

With an annual intake of between 180 and 240 students, we had to create 20 to 28 PBL groups and needed many tutors to be involved in running such small group learning activities. The involvement of many tutors facilitating and assessing students in the various groups over an academic session may result in large inter-rater variability and consequently a less meaningful assessment. Therefore, the development of an assessment instrument, including its related detailed criteria guidelines, is necessary in order to focus tutors' attention on specific competencies to be assessed and to reduce subjectivity of scoring. The assessment of student performance during the PBL tutorial discussions also contributes to 5% of the total mark of the course. Although the instrument is used for summative assessment purposes, it may also function as a formative assessment since it helps tutors give valuable feedback[12] to students with regard to their performance and skills development.

This paper reports (1) the development and evaluation of a summative assessment instrument for measuring student performance in the classroom context, including its "acceptability" and "feasibility" of use by the tutors, and (2) the evaluation of consistency of ratings or assessment scores by different raters on the students' performance in their PBL tutorial group discussions, using the instrument that we have constructed.

Materials and Methods

Development of the Assessment Instrument

An assessment instrument that addresses 4 areas of competency was designed by 3 of the authors (SMS,

Table 1. PBL Tutorial Assessment Form Used for Assessing Student Performance During PBL Tutorial Discussion

UNIVERSITY OF MALAYA
MBBS Year __ - Session 200_/200_
PBL Tutorial Assessment

Scenario: ________________________ (Session: __)
Lab Group: __________

		Performance *(Please circle 1, 2, 3, 4 or 5)*				
No.	**Students' Names**	**Participation and Communication Skills**	**Cooperation / Team-building Skills**	**Comprehension / Reasoning Skills**	**Knowledge / Information gathering Skills**	**Other Remarks**
1.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	
2.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	
3.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	
4.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	
5.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	
6.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	
7.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	
8.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	
9.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	
10.		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	

N.B.: 1 = Unsatisfactory 2 = Marginal 3 = Satisfactory 4 = Good 5 = Outstanding

Signed:

Tutor's Name: ____________________

NMAA and CPLT) with valuable input from a visiting external examiner, who is a faculty of the University of New Mexico, Albuquerque, USA. The assessment instrument was presented in a tabulated form (Table 1), where the tutor was required to rate or score each student's performance in (i) participation and communication skills, (ii) cooperation or team-building skills, (iii) comprehension or reasoning skills, and (iv) knowledge or information-gathering skills. A 5-point grading system was used where the score of 1, 2, 3, 4 or 5 represents "unsatisfactory", "marginal", "satisfactory", "good" or "outstanding" performance, respectively. Scores for the 4 areas of competency were averaged over 2 sessions to give a score of 1 to 5 for each student on each case. Absentees were given a score of zero for the purpose of summative assessment. However in this study, their data were computed to produce intrapolated scores, which were subsequently used to estimate the intraclass correlation coefficients. An appended sheet (Table 2) detailing criteria guidelines for scoring the 4 areas of competency mentioned earlier was also provided to each tutor.

Acceptability of the Assessment Instrument

Tutors who had used the instrument were surveyed to determine whether the assessment instrument developed was user-friendly and the criteria guidelines provided were helpful. This tutor survey was carried out towards the end of the 2001/2002 academic session. Only tutors who had used the assessment instrument or form for at least 2 PBL cases (i.e., who had used this instrument at least 4 times) prior to the study were invited to participate in this survey. They were asked to provide a score, using a 5-point Likert

Table 2. Criteria Guidelines for the Assessment of Student Performance in PBL Tutorials

Score	**Participation and Communication Skills**	**Cooperation / Team-building Skills**	**Comprehension / Reasoning Skills**	**Knowledge / Information Gathering Skills**
1	• Does not respond to verbal / non-verbal cues from others • Does not speak or listen to others or only to tutor·	• Does not contribute to identifying learning issues • Does not give others the opportunity to speak or interrupts others • Unwilling to acknowledge others' views or take up any task	• Does not demonstrate understanding of basic (biological, behavioural and/or population) concepts • Does not seek clarification of concepts	• Has no recall of previous knowledge • Not prepared for session
2	• Rarely asks questions • Responds only to verbal cues • Shows limited non-verbal response during discussion • Discussion or description cannot be understood by others	• Rarely participates in identifying the learning issues • Takes up task only when asked to by the others • Tends to dominate discussion	• Demonstrates understanding of basic concepts with considerable guidance • Rarely seeks clarification of concepts	• Has limited recall of previous knowledge • Prepared for only certain learning issues
3	• Occasionally asks questions • Responds to verbal / non-verbal cues • Occasionally presents ideas clearly	• Volunteers to perform tasks (e.g. to scribe, read case) • Participates in identifying most learning issues	• Demonstrates understanding of concepts with little guidance • Draws reasonable conclusions from given data or information • Often seeks clarification of concepts	•Applies previous knowledge to current issues • Prepared for most learning issues
4	• Regularly asks questions that stimulate discussion • Often presents ideas clearly and helps clarify ideas from others and for others	• Participates regularly in identifying and helps to prioritise learning issues • Encourages others to participate	• Understanding of concepts is demonstrated clearly • Draws valid conclusions with proper interpretation of data or information • Recognises flaws in data or reasoning if pointed out by someone else	• Well prepared for session • Provides references for given information • Recognises integration of knowledge when explained by others
5	• Leads discussion among group members • Constantly presents clear ideas with demonstration of listening, summarising and clarification skills	• Asks for feedback from the group • Organises the group • Shows empathy • Tries to bring quiet members into discussion in a diplomatic manner	• Demonstrates understanding by applying and linking concepts to problems. Explains concepts to others clearly • Integrates difficult concepts • Identifies flaws in data or reasoning independently	• Well prepared for session and identifies key references • Regularly integrates biological with behavioural and population perspectives, providing explanations

1 = Unsatisfactory 2 = Marginal 3 = Satisfactory 4 = Good 5 = Outstanding

© Faculty of Medicine PBL Committee, University of Malaya, 2001

Table 3. Tutors' Responses* on the Use of the PBL Tutorial Assessment Form

No.	Question	Year I Tutors (n=17)	Year II Tutors (n=17)
	The Assessment Form	Mean ± SD	Mean ± SD
1.	In general, I have no difficulty in using the assessment form to assess the students	3.47 ± 1.07	3.76 ± 0.83
2.	I believe I have assessed the students fairly using the assessment form	3.41 ± 1.23	3.71 ± 0.77
3.	The 5 point (1-5) grading system is better than a 3 point (1-3) grading system.	3.65 ± 1.27	4.06 ± 0.97
4.	The criteria given for the grading system are helpful	3.53 ± 1.07	3.53 ± 1.37

SD: standard deviation

*Responses were based on 5-point Likert scale, where *1 = strongly disagree, 2 = agree, 3 = somewhat agree, 4 =disagree, 5 = strongly disagree*

scale (1 = strongly disagree, 2 = disagree, 3 = somewhat agree, 4 = agree, and 5 = strongly agree), on 4 questions related to the use of the assessment form, as part of a larger general survey on PBL-related activities. The questions are shown in Table 3.

Reliability of the Assessment Instrument

In the 2002/2003 academic session, the assessment instrument was refined based on feedback obtained from tutors in the 2002 survey, and was then formally adopted for use in the PBL tutorial sessions to score students' performance during their tutorial discussion. Commencing from 2002/2003, the accumulated scores obtained by each undergraduate medical student contributed to 5% of his or her total summative assessment scores for that academic session. To determine whether this assessment instrument was reliable, the assessment scores of students in various PBL groups in Year I and Year II were compiled and then analysed retrospectively.

Although data have been compiled for at least 2 academic sessions for each pre-clinical year, we describe here a study that evaluated the performance of one cohort (2003 to 2008) of students when they were in Year I in 2003/2004 and Year II in 2004/2005. The Year I students studied only 3 PBL cases, while the Year II students studied 8 PBL cases. Each case was discussed over 2 PBL tutorial sessions (3 hours per session for Year I and 2 hours per session for Year II), scheduled about 1 week apart. There were 28 tutorial groups in Year I with 8 to 9 students in each group and the participation of a total of 27 Year I tutors. In Year II, this same cohort of students was rearranged into 24 new tutorial groups with 9 to 10 students in each group, involving a total of 65 Year II tutors. Unlike the conventional PBL practice, our tutors were not attached to a particular PBL group but took turns to facilitate different groups for different case or problem discussions. Tutors were provided a set of criteria guidelines for scoring students' performance in these 4 areas of competency. Rating scores given by each PBL tutor for the 4 specified areas of competency over the 2 PBL tutorial sessions were averaged to give a final mean score of 1 to 5 per case for each student in the Year I and Year II PBL groups.

Statistical Analysis

Data were reported as mean ± SD of n items. The data on students' scores were analysed using repeated measures ANOVA (Statistical Package for the Social Sciences, SPSS, version 13; SPSS Inc, Chicago, Illinois, USA). The reproducibility of the assessment scores given by the various tutors to individual students of a PBL tutorial group was estimated using the intraclass correlation coefficient (ICC), which was calculated using the formula

$$\frac{MS_{(subject)} - MS_{(error)}}{MS_{(subject)} + (k-1)MS_{(error)}}$$, where MS = Mean Square,

and k = number of measurements, which in this case is the number of scores assigned by different tutors to a student. This ICC formula was transformed computationally from the reliability coefficient formula

$$\frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}$$ where σ_s^2 = between-subjects variance

and σ_e^2 = error variance, sometimes also known as residual or within-subjects variance. The between-subjects variance, sometimes known simply as subject variance, is related to the Mean Square (Subject). The error variance is variation for which we have no ready explanation, and is related to Mean Square (Subject x Case). Thus, ICC is the proportion of variance in the scores related to true variance between the objects of measurement, often called "subjects" – in this case, students. The difference between 2 means was tested using unpaired student's *t*-test and a *P* value of <0.05 was considered statistically significant.

Results

Acceptability of the assessment instrument

A total of 17 Year I tutors and 17 Year II tutors responded to the survey conducted in 2002. The mean ± SD scores of these tutors' responses to the 4 questions posed to them in the survey are presented in Table 3. The majority of the tutors surveyed agreed that they had no difficulty in using the instrument (82.4% Year I and 94.1% Year II tutors scored ≥3), that they had assessed the students fairly using this instrument (70.6% Year I and 88.2% Year II tutors scored ≥3), and that the criteria guidelines were helpful in assessment scoring (76.5% Year I and 88.2% Year II tutors scored ≥3). They also preferred a 5-point to a 3-point grading system (82.4% Year I and 94.1% Year II tutors scored ≥3).

Reliability of the Assessment Instrument

Graphs showing the scores given to students in several representative groups, one each of Year I and Year II, by their respective tutors are presented in Figures 1 to 4. These graphical data showed that in some groups (Fig. 1), there were obvious differences in the performance of individual students (e.g., the mean score per case ranged from 2.1 ± 0.6 to 4.1 ± 0.3 in the Year I group, n = 3 cases; and from 2.8 ± 0.8 to 4.4 ± 0.4 in the Year II group, n = 8 cases). These are examples of groups with heterogeneous composition of students where a >1.5-fold difference was observed between the highest and lowest mean scores (*P* <0.05).

On the other hand, there were groups where the individual students' scores, given by the different tutors, were consistently similar to each other (Fig. 2). For example, the scores for the various group members in a Year I group ranged between 3.5 ± 0.5 and 4.2 ± 0.2, and in a Year II group, ranged between 3.6 ± 0.8 and 4.2 ± 0.4. These are examples of groups with homogeneous composition where the difference between the highest and the lowest mean scores of the group members was <1.2 times and the difference was not statistically significant.

In addition, this study showed that a small number of tutors consistently showed "strict" (2 each in Years I and II) or "indiscriminate" rating practice (3 in Year I and 5 in Year II) in the academic session studied (Figs. 3 and 4, respectively). The "indiscriminate" raters also tended to give higher scores of 4 to 5 for each student. The reasons given by some of these "indiscriminate" tutors for such practices included unwillingness to comply with the guidelines, a desire to encourage all the students in the group, and a score supposedly reflective of the group's collective effort. When the number of tutors or cases involved was larger (as in the Year II groups), the change in the mean scores caused by these "strict" or "indiscriminate" raters was relatively small (compare the top with the bottom panels of Figures 3 and 4).

ICCs were calculated only for the Year II groups since the Year I tutor group sample size (n = 3) was too small. The estimated ICC values for the 24 Year II groups (including the "strict" and "indiscriminate" raters) ranged from 0.16 to 0.75 (0.42 ± 0.16, n = 24), indicating a vast range of reproducibility in the scoring by tutors in the different groups. When the scores given by these "strict" and

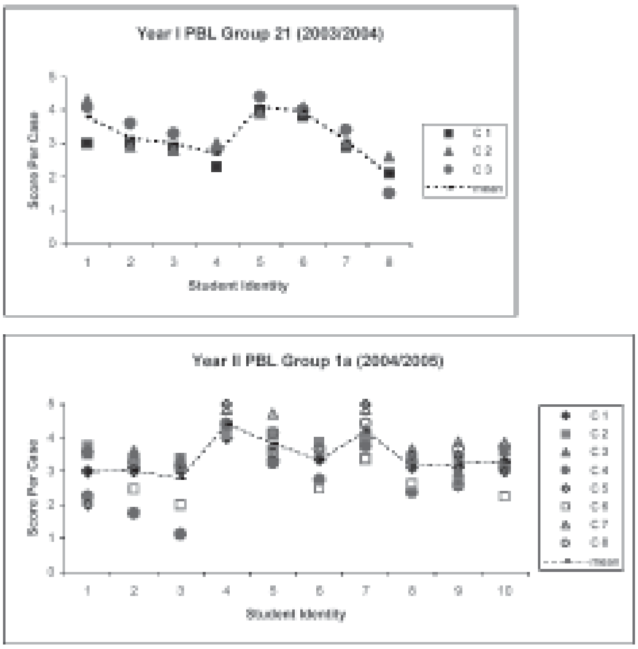

Fig. 1. The scores obtained by individual students of a representative PBL group each in Year I (Top panel: Group 21, 8 students, 3 cases) and in Year II (Bottom panel: Group 1a, 10 students, 8 cases), given by different raters for the cases studied. The dotted line represents the mean scores obtained by individual students of each group for all the cases studied. These are examples of groups with heterogeneous composition.

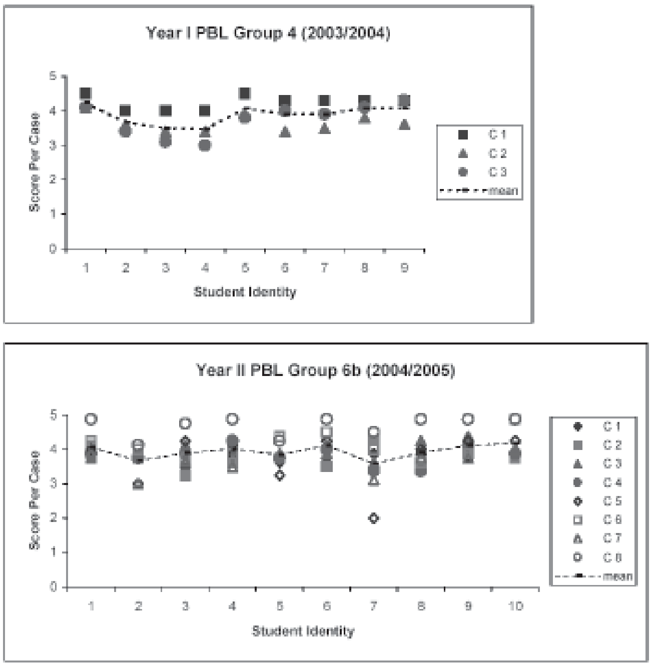

Fig. 2. The scores obtained by individual students of a representative PBL group each in Year I (Top panel: Group 4, 9 students, 3 cases) and in Year II (Bottom panel: Group 6b, 10 students, 8 cases), given by different raters for the cases studied. The dotted line represents the mean scores obtained by individual students of each group for all the cases studied. These are examples of groups with homogeneous composition.

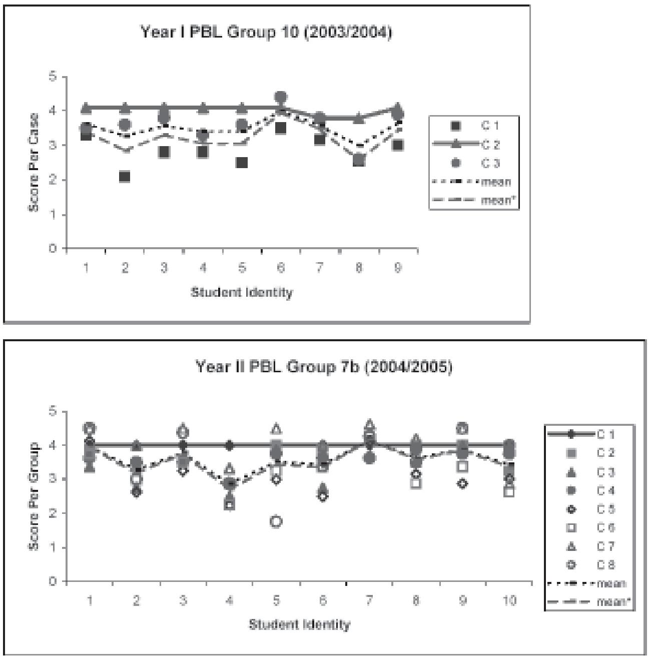

Fig. 3. The scores obtained by individual students of a representative PBL group each in Year I (Top panel: Group 10, 9 students, 3 cases) and in Year II (Bottom panel: Group 7b, 10 students, 8 cases), given by different raters for the cases studied.

The solid line represents the scores given by an "indiscriminate" rater. The dotted lines represents the mean scores obtained by individual students of each group given by all raters, including (**.........**) or excluding (– – –) the "indiscriminate" rater.

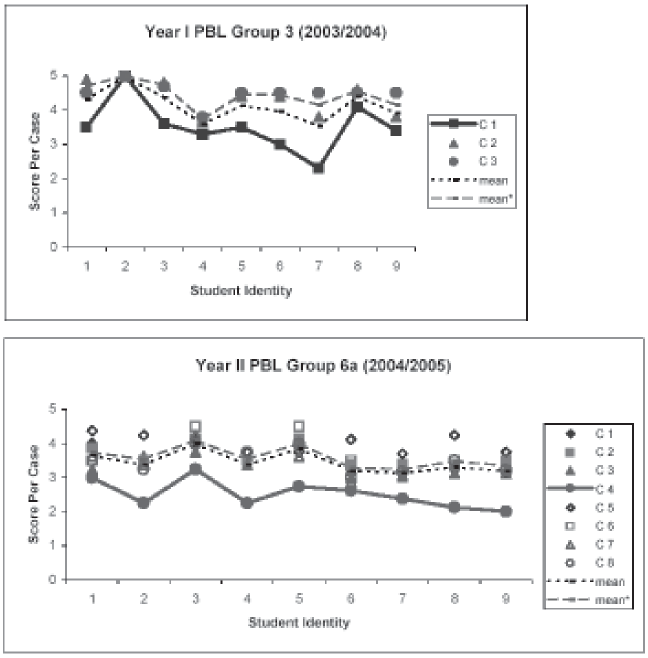

Fig. 4. The scores obtained by individual students of a representative PBL group each in Year I (Top panel: Group 3, 9 students, 3 cases) and in Year II (Bottom panel: Group 6a, 9 students, 8 cases), given by different raters for the cases studied.

The solid line represents the scores given by a "strict" rater. The dotted lines represents the mean scores obtained by individual students of each group given by all raters, including (**.........**) or excluding (– – –) the "strict" rater.

"indiscriminate" raters were excluded during the computation, the reproducibility of scores improved (the mean ICC for the 24 groups increased to 0.48 ± 0.16).

Discussion

A review on professionalism in medicine[13] defines medical professionalism as the ability to meet the relationship-centred expectations required to practice medicine competently. This relationship varies along a continuum from positive to negative engagement and encompasses constructs such as respect for others' integrity and diversity. To achieve such a positive relationship, appropriate communication and interpersonal skills (C-IP skills) are required.[14] It has been documented in the literature that excellent interviewing skills can strengthen the bond between a patient and his or her doctor. It has also been stated that effective C-IP skills increase patient satisfaction and are associated with patient compliance, improved health status and resolution of symptoms.[14] When considering the group dynamics of students in the PBL tutorials, these C-IP skills are also important for effective discussion. Furthermore, such skills are often utilised and unconsciously practised by the students during these tutorial sessions. Since assessment drives learning, the assessment of students undergoing these activities would induce and encourage them to improve these C-IP skills. However, assessments of the performance of medical students in a classroom setting, such as during a PBL tutorial discussion (or a *viva voce*), are examples of assessments that depend on the consistency of raters and their ratings for reproducibility and reliability. The largest threat to the reproducibility of such ratings has been shown to be rater inconsistency and low rater or inter-rater reproducibility.[15]

In this study, analysis of data compiled for the cohort 2003 to 2008 indicates that the different raters (a total of 3 for each Year I and 8 for each Year II group) were relatively consistent in their assessment of students' performance within the group, despite there being a few "strict" or "indiscriminate" raters. Graphic displays of the scores for the same group of students rated by the different tutors generally show a similar scoring profile within a PBL group. This indicates a consistency of judgement with respect to the relative performance of the students within the group.

ICC has been used as a method for estimating inter-rater reliability. In this study, a large range was obtained for the calculated ICC values. Possible contributing factors to the apparently large range of calculated ICC values or low intraclass correlations of the assessment scores include the following:

1. The error variance was affected by both variations in the raters involved and cases studied. This happened because different tutors rated the same group of students at different time using different cases studied.
2. For groups with a very homogeneous composition of group members (e.g., Group 6b of Year II), the between-subjects variance would be relatively small (e.g., σ_s^2 = 0.02 for Group 6b) compared to the corresponding error variance (e.g., σ_e^2 = 0.12) leading to a low ICC value (0.16). In contrast, for groups with a heterogeneous composition of members (e.g., Group 1a of Year II), the reverse tends to be true, where the between-subjects variance is larger (e.g., 0.28 for Group 1a) compared to the corresponding error variance (0.12), leading to a large ICC value (0.70).
3. The practice of "indiscriminate" scoring by some raters in assigning the same scores for all may mask the true variability among the group members as observed by the other raters. This effect is expected to be greater with a heterogeneous than with a homogeneous group of students.
4. The practice of "strict" scoring by some raters has also increased the range of inter-rater scores, indirectly influencing the error variance, even though the relative position of the group members' scores were not affected.
5. Unwillingness of some raters to follow the criteria guidelines, or marked differences in their interpretation of the criteria provided, would affect both the between-subjects and error variance values.

In light of the above observations, it would be useful to apply this assessment instrument in a setting where different tutors rate the performance of a PBL group of students on the same case. This is expected to give a more accurate reflection of the inter-rater variability, and thus the reliability of the instrument. On the other hand, this instrument may also be used to assess inter-case variability if the same tutor rate the same group of students for all the cases studied.

Notwithstanding the above shortcomings, our analysis on the use of this process-oriented instrument suggests that the use of a rating form is actually an effective way to evaluate the level of C-IP skills since it provides a hierarchy of responses to indicate how well the student performed in the evaluation. The introduction of the criteria guidelines helps to reduce the effect of inter-rater variability (as evidenced in an unpublished observation and also part of the 2002 survey) for the majority of our tutors. There were, however, a few tutors who did not adhere to the given guidelines or used "strict" criteria than the others. In such cases, the increase in the number of raters for each of the PBL tutorial groups would help to reduce their effect. Omission of the scores given by these "non-compliant" raters showed that the average scores attained by the students only "shifted" slightly (as shown in Figure 1, bottom panel). While we cannot compel every tutor to follow strictly the guidelines for rating, some measures have been taken to reduce "non-compliance" among tutors. They include reminding tutors of the purpose of the tutorial assessment, clarifying queries on the use of the guidelines for rating, providing feedback on the findings of our inter-rater variability, and highlighting to those "indiscriminate" or "strict" raters the deviation of their rating from the majority of other tutors.

Conclusion

The instrument designed for the assessment of student performance in the PBL tutorial classroom setting, when used judiciously with the criteria guidelines provided, is feasible and reasonably reliable. While there is a reasonable amount of consistency in the raters' judgement of students' performance in the PBL tutorial, further training in the use of the assessment instrument and understanding of its purpose is likely to improve the inter-rater consistency of assessment scores.

Acknowledgements

The authors would like to thank Professor Geoff Norman and Dr KF Quek for their advice on statistical analysis, and Azlila Yunus, Noor Shafila Shafie and Kamila Banu Rahim Khan for helping us with data entry. The authors would also like to thank the Foundation for the Advancement of International Medical Education and Research (FAIMER) for their academic support and encouragement for two of the authors (NMAA and CPLT) who are fellows of the FAIMER Institute Programme.

This study was funded by the Intensive Research in Priority Areas (IRPA) grant (07-02-03-0227), from the Ministry of Science, Technology and Innovations (MOSTI), Malaysia.

REFERENCES

1. Newble D, Jaeger K. The effect of assessment and examinations on the learning of medical students. Med Educ 1983;33:165-71.
2. Waas V, Van der Vleuten CPM, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001;357:945-9.
3. Harden RM. How to assess students: an overview. Med Teacher 1979;1:65-70.
4. Cohen-Schotanus J. Student assessment and examination rules. Med Teacher 1999;21:318-21.
5. Azila NMA, Atiya AS, Alhady SF, Tan NH, Sim SM, Nah SH, et al. Adaptive changes towards problem based learning in the New Integrated Curriculum (NIC) of the Medical Course at the University of Malaya. In: Marsh J, editor. Implementing Problem-Based Learning. Proceedings from the 1st Asia-Pacific Conference on Problem-Based Learning. The Problem-Based Learning Project. Hong Kong University 2000:323-31.

6. Azila NMA, Sim SM, Atiya AS. Encouraging learning how to fish: An uphill but worthwhile battle. Ann Acad Med Singapore 2001;30:375-8.
7. Azila NMA, Sim SM. The status of problem-based learning in the medical schools in Malaysia. J Med Educ 2005;9:121-30.
8. Schmidt HG, Moust JHC. Factors affecting small-group tutorial learning: a review research. In: Problem-based Learning: A Research Perspective on Interactions. Mahwah, NJ: Lawrence Erlbaum, 2000:19-52.
9. Azila NMA, Tan NH, Sim SM, Rosnah I, Liam CK, Atiya AS. Critical thinking development in a "hybrid" problem-based learning programme. J Med Educ 2003;9:306-12.
10. Barrows HS. Practise-based Learning: Problem-based Learning Applied to Medical Education. Springfield, Illinois, Southern Illinois: Southern Illinois University School of Medicine, 1994.
11. Neville A. Student evaluation in problem-based learning. Pedagogue 1995;5:2-7.
12. Rolfe I, McPhearson J. Formative assessment: How am I doing? Lancet 1995;345:837-9.
13. Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: a review of the literature. Med Teacher 2004;26:366-73.
14. Hobgood CD, Riviello RJ, Jouriles N, Hamilton G. Assessment of communication and interpersonal skills competencies. Acad Emerg Med 2002;9:1257-69.
15. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004;38:1006-12.