Peer assessment of professional behaviours in problem-based learning groups

Chris Roberts

Christine Jorm

Jim Crossley

Stacey Gentilcore

Abstract

Background

It is a common conception that peer assessment of professional behaviours within small group activities such as problem-based learning (PBL) can be valid and reliable. Consequently poor student scores may lead to referral to faculty based student support or disciplinary systems. We wished to determine whether a multisource feedback tool measuring professional learning behaviours in PBL groups is sufficiently valid for decision-making about student professional behaviours.

Methods

Data were available for two cohorts of students who were learning in PBL groups. Each student was rated by his or her PBL group peers on a modified version of a previously validated professional learning behavior scale. Following provision of feedback to the students, their behaviours were further peer assessed. A generalisability study was undertaken to calculate the students' professional behaviours, sources of error that impacted the reliability of the assessment, changes in rating behaviour, and changes in mean scores after receiving feedback.

Results

A peer assessment of 'professional' learning behaviour within a PBL groups was highly reliable for 'within group' comparison (G = 0.81-0.87) but poor for 'across group' comparison (G= 0.47 - 0.53). This was because the stringency of fellow students as assessors was so variable and they are nested within groups. Feedback increased the range of ratings given by assessors and brought their mean ratings into closer alignment. More of the increased variance was attributable to assessee performance rather than assessor stringency, so there was a slight improvement in reliability, especially for comparisons across groups. Professional behaviour scores were unchanged.

Conclusion

A multisource feedback tool measuring professional learning behaviours in PBL groups is unreliable for decision-making outside a PBL group. Faculty should not draw any conclusions from the peer assessment about a students' behaviour compared with their peers in the cohort. The provision of a summary of the peer feedback had a demonstrable effect on students' behaviour as peer assessors, by providing formative feedback on their own behaviour from their PBL group peers, but not on their own professional behaviour. Health professional educators need to reframe the question of assessing professional behaviours in PBL groups to focus on opportunities for formative peer feedback and its impact on learning.

Keywords. Peer Assessment, Problem Based Learning, Generalisability, Professional Behaviour.

Background

There are compelling reasons why medical students need to learn how to both give and receive feedback. Feedback is a critical component of fitness to practice. Doctors are known to be reluctant to report incompetent and impaired colleagues, [1, 2] yet the need for reporting often arises after a sustained failure to give effective peer feedback. In the UK, colleague and patient feedback is one element of the supporting information that the medical council requires doctors to collect and reflect upon as part of the process of revalidation. [3] In Australia, being honest, objective and constructive when assessing the performance of colleagues, including students, is part of good medical practice. [4] Multisource feedback for doctors (also known as '360 degree' feedback) with comment from peers, supervisors, and other health professionals is becoming a common method of collecting this kind of feedback in the work-based assessment of junior doctors in many parts of the world. [5 6, 7 8 9]MSF provides an efficient, questionnaire-based assessment method that provides feedback about clinical and non-clinical performance to trainees across specialties and is considered valuable for both formative and summative assessments. [7] It is thought to lead to performance improvement, [10] although individual factors, the context of the feedback, and the presence of facilitation have a profound effect on the response. More broadly, resistance to accepting feedback is considered a marker of unprofessional behaviour. [11]

Some medical educators have incorporated the assessment of professional behaviour into the medical school curriculum to offer the opportunity of early detection and timely remediation for students who exhibit dysfunctional behaviour. Such behaviours are assessable by a variety of methods. [12] Some medical schools provide a general reporting facility. [13] Yet students struggle with reporting an unprofessional peer lest they bring harm to the peer, themselves, or the group they are working in. [14] Students are often reluctant to give feedback on peers.[15] On the other hand, students are seen as having a key role in driving learning, and thus should be generating and soliciting their own feedback. [16] Feedback may demonstrate what is understood by ''good'' behaviour, and help diagnose the gap between a student's current behaviour and the desired behaviour. [17]

Methods such as peer 360-degree feedback for assessing undergraduate medical students' personal and professional behaviours are thought to have sufficient utility to be used summatively despite student ambivalence towards judgments of these behaviours. [18] They are thought to have high reliability and to provide stable estimates of error variance across independent cohorts of raters, [19] if a sufficient number of observers are used. [20]

Some schools have made use of problem-based learning (PBL) as an opportunity for peer assessment of a range of professional behaviours. [14, 20 21 22] A learning environment such as that of PBL, with its characteristic features of students working together in small group in a highly self-directed and experiential learning activity, seems to be eminently suited to fostering appropriate professional behaviours. [23] Whilst students may be concerned about peer assessment, [13 20, 24] the quality of the contributions that students make during tutorials strongly affects the quality of the discussion and therefore group functioning.[22] Whilst tutors have only limited time to observe each student, students have many opportunities to observe each other.[13, 25] A number of studies have measured tutorial group effectiveness based on student self-assessment of particular behaviours [26 27.28] However, the literature is equivocal as to whether such assessments are optimal in terms of reliability and validity. [22] Methods of constructing and delivering constructive and objective peer feedback can be taught.[29] For example, several years experience with peer assessment in Rochester School of Medicine has demonstrated that peers can provide reliable, stable ratings of both work habits (e.g. preparation, problem solving and initiative) and interpersonal attributes (e.g. truthfulness, respect, integrity and empathy).[25]

There is a scarcity of research investigating the validity of peer assessment tools within PBL. [30] So far, research using a recognizable framework of validation [31] of the peer assessment of professional behaviours in the PBL tutorial groups has focused on establishing the internal structure of the assessment and its relationship with other variables of interest. Kamp *et al*., (2011) developed and validated the Maastricht-Peer Activity Rating Scale (M-PARS), a tool measuring constructive, motivational, and collaborative factors, in the PBL tutorial. [22] It was found to have a good model fit using a confirmatory factor analysis, with high correlations between the three subscales. In addition, generalizability studies were conducted in order to examine how many different peer ratings per individual student were necessary to ensure a reliable evaluation of one student. When students were evaluated by, at least, four of their peers, the G co-efficient was 0.77. However, the design of the G study was not reported, and it was unclear whether this figure related to student's peer assessment scores within their own PBL group, or students' scores across all PBL groups. Van Mook et *al*., [32] have shown that web-based peer assessment of professional behaviours is acceptable to students and significantly increased the amount although not the quality of written feedback, suggesting that five raters was optimal. Papinczak *et al.,* (2007) developed a peer assessment instrument, in which internal consistency (Cronbach's alpha) of peer averaged scores across all PBL groups ranged from 0.66 – 0.77. Other evidence of validity were reported demonstrating peer averaged scores correlating moderately with tutor ratings initially (r = 0.40) and improving over time

(r = 0.60). Students consistently over-marked their peers, particularly those with sceptical attitudes to the peer-assessment process. Generalisability theory was not used. [33] Papinczak *et al*., (2007) also demonstrated correlations with both self- and Faculty-based assessments in the PBL tutorial, and showed modest correlations. Reiter *et al*., (2002) developed an instrument for students in PBL groups to rank the professional behaviours of their peers, [34] which proved an unreliable measure of tutorial performance, because ratings were inconsistent from one week to the next as well as across raters within a week. Sullivan *et al*., (1999) found a moderate correlation between peer and tutor ratings, and very little correlation between self- and tutor ratings. [35] There is thus a need for more validation research [31] in the context of peer assessment within PBL.

Theoretical framework

An opportunity to further explore the validity of peer measures [31] of professional behaviours in PBL arose when Sydney Medical School implemented a peer assessment instrument, where students rated their peers in their PBL tutorial groups as part of the required formative assessment of the personal and professional development theme. Approximately two weeks after the first peer assessment at the end of a PBL block, all students received an individualised summary of the peer feedback that their PBL group members had provided. A second peer assessment was undertaken in a later block. We detail this intervention in the methods section. We were interested to identify four aspects in the validity argument, derived from Kane [31]: *Scoring* (empirical evaluation of the multi-source feedback instrument); *Generalisation* (using the observed scores to generate an overall test score representing professional behaviour in the PBL tutorial setting); *Extrapolation* (drawing an inference regarding what the test score might imply for the professional behaviours of students), and *Implications* (if the scores were credible and reasonably free from error, could they be used as a summative assessment of professional behaviour and provide an opportunity of early detection and timely remediation for students who exhibit dysfunctional behaviour). In this paper we focused on the generalisation argument and its implications. We wished to investigate empirical evidence on the generalizability of student assessor scores. In particular we were interested to derive the sources of error in the peer measurement, related to assessor subjectivity, which did not relate to the construct of interest. These include assessor stringency/leniency, which is a first-order effect and is defined as the consistent tendency of assessors to use either the top or the bottom end of the rating scale. Assessor subjectivity refers to assessor preference for assessee and includes how different assessors favour different examples of questions differently over and above their baseline stringency. [36, 37]

In this paper, we posed three research questions: a) What is the degree of student assessor stringency/leniency and subjectivity in a peer assessment of professional behaviours within a single PBL group and across PBL groups? b) What is the impact on student assessor stringency/leniency and subjectivity of receiving feedback on how he/she was scored by other members of their PBL group? c) To what extent are changes in student's peer assessment scores after receiving feedback related to changed rating behaviors and to what extent are they related to changed professional behaviours.

Methods

Instrument development

The feedback instrument in this research was intended to stimulate reflection with two purposes in mind: (1) to cue desirable behaviour changes as the student considered their own performance profile and (2) to cue desirable rating behaviour changes as the student considered experientially, from a recipient's perspective, the meaning and significance of the rating responses that they are providing for others. This instrument was a modified version of a previously validated peer assessment instrument for use in PBL groups. [21] The original scale had been developed by Papinczak et al., (2007) using quantitative and qualitative data collected from tutor assessment of students' PBL performance at the University of Queensland. The original tool included seventeen items across five domains; responsibility and respect, information processing communication, critical analysis, and self-awareness through presentation of a case summary. Students rated their strength of their agreement or disagreement with the statements about their peers' performance in that week of PBL tutorials using a five point Likert Scale (1 = totally disagree; 5 = totally agree). It was designed to be used after the designated student had presented a summary to the group, which gave the student a specific communicative and educational leadership opportunity. The version of the Papinczak et al., (2007) instrument used in this research was further modified following consultation with University of Sydney PBL tutors, who for pragmatic reasons recommended a shorter version of the scale. They were concerned that the long version would take too much time. The final scale used in this research consisted of 9 items across 5 domains, and a global rating (item 10) and is given in Figure 1. The mean score of the instrument for each student was calculated across the nine items and the ratings of students within their PBL group (n = 9-14).

Insert figure 1 about here

PBL in the Research Context

Sydney Medical School had introduced PBL in 1997 within a four-year graduate entry program and originally used a three tutorial-based system in the first two years of the course. This was delivered by an integrated IT system to control and manage content, and the program has been extensively evaluated. [38, 39].[40] However in line with many medical schools internationally, consideration of the resources needed to sustain PBL [41] saw the adoption of a two tutorial system in 2012. Working in collaboration with group members, students analyse a problem of practice, formulate hypotheses, and undertake self-directed learning to try to understand and explain all aspects of the PBL problem. The explanations are encouraged to be in the form of an underlying process, principle, or mechanism. The two 1.5 hour tutorials are held on the same day, with the first being student-led using the extensive IT materials to support the development of the case and tutor facilitation being provided for the second tutorial. Each block is eight weeks long. Students receive an orientation to the format of the PBL at the start of the first semester and can access a student handbook detailing the instructional method. Students are randomised into PBL groups, which then change as students move from first year to second year in their program. Students are expected to attend all PBL sessions.

Peer assessment

The quasi-experimental design used to answer our research questions is given in Figure 2. The peer assessment of PBL performance was made a required formative assessment within the Professional Development (PPD) curriculum theme, and thus participation for students was compulsory. Students had received prior formal instruction on best practice in feedback. Each student was invited to complete the on-line assessment for every other student in his or her PBL group within one week of the invite. Additionally each student was required to also give constructive written feedback on the contribution to the PBL group of 4 group members. The software assigned two of the student names and the student giving feedback could choose two. Thus each student in a year group assessed themselves and was assessed by his 9 (up to 13) PBL group peers on the professional learning behavior scale consisting of 9 checklist items rated 1-5 (5 being good) and a global rating. This was done on two occasions approximately 20 weeks apart. For the 1st years, this was in blocks two (musculoskeletal) and five (cardiovascular), and for the 2nd years blocks seven (endocrine) and nine (gastroenterology). There was no tutor feedback collected in this period.

Insert Figure 2 about here

Feedback

About 2 weeks after the completion of the block (block two for the 1st year students, and block five for the 2nd years), students could access a confidential report on-line that summarised his or her feedback from their peers. A sample of the quantitative report is given in Figure 3.

Insert figure 3 about here

For each checklist item in the scale, the student received their self-rating and the averaged students' score from the year cohort. They also received the anonymised free text comments that their peers in the PBL group had made. It has been reported that there has been some degradation of the PBL process in other settings, [28 42] resulting in dysfunctional behaviour. This was a concern in our setting because of the large group size and the student led first PBL session. Receiving this type of feedback was anticipated to change their professional behaviours, and see an increase in their mean score on the second occasion of peer assessment. The students' rating behaviour was expected to improve by reducing assessor subjectivity. The PPD Coordinator reviewed students with low scores in the peer assessments. The written feedback is not considered within this paper.

Data Analysis

In Generalisability theory [43] the G–study provides a means of quantifying the sources of potential error in the assessment simultaneously, using all of the available data. The student's universe score consists of all the trials of the assessment design that might hypothetically be carried out, using innumerable sets of tasks, administered on distinct occasions, with innumerable scorings of each performance by informed assessors.

A variance components analysis estimated the contribution that the wanted factor (the professional behaviour of the student) and the unwanted factors (e.g. the impact of the assessor) made to the variation in peer assessment scores. Variance estimates were then combined [36] to provide an index of reliability (the G coefficient). The strength of this approach is that future modifications of the assessment program can be planned that address the main sources of error identified in the initial study.

We used the General Linear Model within SPSS (version 20) to undertake a G study. The overall checklist score was used as the dependent variable because factor analysis demonstrated a unitary

structure and the items had been chosen to provide an exhaustive representation (rather than a sample) of professional behaviour in the PBL context. The G-study was based on a nested model with students as assessees (p) and students as assessors (j) both nested within PBL groups. D-studies were conducted to model the reliability of two scenarios. The reliability of comparisons between students within a PBL group is given by:

$$G=Var_p/(Var_p+Var_{pj})$$

The reliability of comparisons between students across groups but within a cohort is given by:

$G=Var_p/(Var_p+Var_j/n+Var_g+Var_{pj}/n)$ – where n is the number of student assessors and g is the PBL group.

A standard error of measurement (SEM) was calculated from the square root of the error variances in the appropriate denominator. 1.96 times the SEM gives a 95% confidence interval for a hypothetical 'typical' student's score. For our first two research questions, we were interested the circumstances in which the variance of students' scores within the PBL group was greater or less than the variance of students' scores across groups. [44] For our third research question, we were interested in whether the mean peer assessment score would change as a result of the summarised feedback given to students.

Ethics

The University of Sydney research ethics committee has a long standing agreement with the Sydney medical school where students on entry sign a waiver to allow the use of their anonymised routine collected assessment data for evaluation and research purposes.

Results

Data on peer assessment rating were available for two separate cohorts within the same academic year on two occasions each. For 1st years, there were 305 students learning in 28 PBL groups of 9 to 12 students. For 2nd years, there were 328 students in 28 PBL groups of 10 to 14 students. The results of the generalizability study are given in Tables 1 and 2.

Insert Table 1 and 2 about here

As expected, both wanted and unwanted facets contribute to the variation in professional behavior score. Across both cohorts and both iterations the largest contributor to variance was variation in assessor stringency (Var_j), followed by assessor subjectivity (Var_{pj}), followed by assessee differences (Var_p). PBL group means (Var_g) also varied slightly. This kind of pattern is not unusual in judgement-based assessment. [45] [46] It means that each individual rating is not reliable because it is determined more by which assessor made the judgement than by which student was being assessed. However, as more judgements are considered, the combined result becomes more and more reflective of the differences between students.

Overall, the 1st Year students produce more variable ratings than the 2nd year students (higher absolute variance estimates) and did so without increasing their stringency variation (Var_j). In plain English, this means that they used more of the scale but without tending to become more 'hawkish' or 'dovish'. Some of the extra variance is attributable to greater assessor subjectivity (Var_{pj}), and some is attributable to assessee differences (Var_p). Of these two Var_p is proportionately greater than Var_{pj}, which means that the 1st year students' ratings provide more reliable comparisons between their peers than the 2nd Year students' ratings. This is particularly true when comparing across groups where the assessors' stringency is nested and causes error.

Across both cohorts exactly the same pattern of differences is seen in the post-feedback data when compared with the pre-feedback data. That is to say the post-feedback data is more like the 1st Year data with greater overall use of the scale without an increase in the variation of assessor stringency/leniency, and most of the extra variance attributable to assessee performance rather than assessor subjectivity. This, again, results in more reliable comparisons between students, especially across groups.

The dependability estimates combining the variance components in table 1 and 2 according to the formulae given earlier are given in Table 3.

Insert Table 3 about here

A peer assessment of 'professional' learning behaviour was highly reliable for 'within group' comparison for 1st year students (G = 0.87), and slightly less so for the 2nd year's (G=0.81). However, reliability was poor for 'across group' comparison both for the 1st years (G= 0.53) and 2nd years (G=0.47). There was a slight increase in reliability for both cohorts after receiving the summarised feedback.

There was no significant difference in mean student scores in the peer assessment of professional

behaviours for either cohort and across both iterations of providing peer feedback (see Figure 4). This suggests the changes in variance of student ratings (Table 1 and 2) are more likely to be due to changed rating behaviour than changed professional behaviours. A hypothetical 'typical' tudent's 'true' score has a 95% chance of lying within 1.96 SEM of his or her measured score. For relative ranking of students across groups, the confidence interval crosses more than three quartiles giving less than 95% confidence that a student in the middle of the top quartile has better behaviours than a student in the middle of the bottom quartile. [36] This is another way of understanding the low reliability co-efficient which synthesises the information about score precision and score spread.

Insert Figure 4 about here

Discussion

Our findings show that a peer assessment of professional learning behaviours designed to be used in PBL groups was highly reliable for 'within group' comparison, but poor for 'across group' comparison. This was because the stringency/leniency of fellow students as assessors is so variable and they are nested within groups. We also found that receiving feedback from peers impacts the process of assessing the behaviour of fellow students. Feedback increased the range of ratings given by assessors and brought their mean ratings into closer alignment. More of the increased variance was attributable to assessee performance than assessor stringency so there was a slight improvement in reliability, especially for comparisons across groups. In our study, there was a difference between first year students and second year students, in that ratings were less reliable in the 2nd year PBL groups. However, it is important to remember that both cohorts were first-time raters and first-time recipients of feedback. There was no significant difference in the average professional behaviour performance following summarised feedback. This makes it more likely that the changes in post-feedback ratings are driven by changes in students' rating behaviour rather than changes in students professional behaviours.

Implications

Our findings in support of the validity argument [31] for using a peer assessment tool in the PBL setting run counter to the findings of others. [22, 35] The empirical evidence available so far has claimed that peer ratings are highly reliable and perhaps the most valid when compared with self- or tutor- based assessments.[20] [35] If the purpose of the peer assessment is to determine whether an individual student has met the expected standards of professional behaviour, then our data suggests it would

be unsafe to generalise a student's score derived from his/her PBL group peers as a meaningful measure of student professional behaviour compared with the cohort. Thus, in the PBL context, it is not possible to draw a reliable inference about the professional behaviours of any students outside his or her group. The implication of our findings is that the professional behaviours score cannot be used as a summative assessment of professional behaviour. However the tool has potential for formative feedback. In our data provision of feedback had the same demonstrable effect on assessor performance in the two separate cohorts. The intervention of giving feedback was a positive influence in changing the assessing behaviour of peer assessors through the experience of being assessed and receiving feedback from fellow peers. This may well be a powerful way to influence rating behaviour in a peer group because it causes assessors to consider both the meaning and the implications of ratings in a direct experiential way. This finding is in contrast to others who found that the quality of individual contributions to the tutorial group does not improve after receiving peer feedback, regardless of whether the group is encouraged to reflect or not. [30]

Figure 4 shows visually the main reason why 1st year ratings were more reliable than second year ratings. 1st years don't provide score with more precision; they are actually less precise, but there is much more performance variance within the cohort so that even less precise scores provide more reliable discrimination between students. [17] Our data showed there was little PBL group effect. This contrasts with findings from another study, where a large group effect was demonstrated, where groups of 3-5 social science students worked on a research design project for 10 weeks, which was assessable by academic staff. Here the group effect was thought to be because groups were self-selected rather than randomised. [47]

The question arises as to which type of quantitative feedback will influence student's performance in future iterations of the assessment. Our data suggests that they might be best provided with the averaged peer students' rating from their PBL group, as well as the qualitative comments. In this study they were provided with averaged ratings from the cohort. We do not have the data to know whether it was the 'within group' qualitative feedback or the 'across groups' quantitative profile that was most influential in producing the effect. However, we recommend using both within group averaged ratings combined with the qualitative feedback.

Similarly, we don't have the data to understand what drives rater error in these ratings, but would like to use a further qualitative study to investigate. Students' metacognitive knowledge about the purpose, and likely outcomes of peer assessment influence their engagement in performing such a

task. Faculty will need to prepare the students in order for them to have more understanding about the meaning and implications of peer assessment. They will need to understand how reflection and goal setting can influence their own professional behaviours [30, 48, 49] as well as their peer rating behaviour. Given the paucity of guidance from the literature, a fresh stating point may be redesigning peer assessments using the perspective that students are experts in assessing the behaviours of their peers in the PBL process, and questionnaire design needs to ensure it is asking the kinds of questions that are important to student assessors. A similar approach has had some success in enhancing reliability in the work based assessment literature. [50]

Whilst the process of giving multi-source feedback may provide formative feedback for the behaviours of student within their group, it is unsafe to draw any conclusions about a student's behaviour compared with students in the rest of the year. The degree of unreliability would also be problematic generating cohort data whereby a student's PBL performance could be related to other aspects of student performance and academic outcomes. Health professional educators need to rethink the value of assessing professional behaviours in PBL groups and are advised to focus on the impact on learning and opportunities for formative feedback. More research is necessary to determine wether our results are generalizable to other settings. Medical educators wishing to introduce peer assessment for professional behaviours should consider combining it with other feedback methods in observing students' professional behaviour.

Strengths and Limitations

As far as we are aware this is the first study to estimate the variances of a peer assessment tool within the PBL tutorial both within and across groups. We acknowledge a number of limitations to the study. The major difference between our results and prior work on peer assessment instrument in the PBL setting, [22, 35] may be down to the differences in research design that were used. Unlike some of the prior research, [20 22, 33-35] the peer ratings were collected for the purpose of a required formative assessment of professional behaviour in the personal and professional development theme. We had no control over the pragmatic version of PBL with the first tutorial being led by a student. Students may have performed haphazardly in the PBL tutorial process, [42] with untended negative impacts on both learning processes and outcomes. [28, 42, 51, 52] The fully validated version of the Papinczak et al., (2007) tool was not used, and the short version used in this study may have led to underestimates of the reliability of the tool. The time period between the first and second peer assessment means that students may have been influenced by factors other than the the summarised feedback. It is known for example that collaborative work in PBL tutorials induce social

cohesion through which students have more understanding about their peers' behaviours. [53, 54] This may be one of a number of confounding factors in our research. The generalisability study made use of naturalistic data, and we had no control over which students were assigned in the PBL group formation. We accept a case could be made for including PBL group variance (Var_g) in the numerator of the "across group" D study design as well as the variance of the student (Var_p). This has the effect of slightly increasing the predicted reliability but not substantially enough to change our conclusions. Further empirical studies on peer assessment are a rich area for further research.

Conclusion

A peer assessment tool measuring student professional learning behaviours in PBL groups is unreliable, and therefore not valid for decision-making outside a PBL group. Faculty should not draw any conclusions from the peer assessment about a students' behaviour compared with their peers in the cohort. The provision of a summary of the peer feedback had a demonstrable effect on students' behaviour as peer assessors, by providing formative feedback on their own behaviour from their PBL group peers. Health professional educators need to reframe the question of assessing professional behaviours in PBL groups to focus on opportunities for formative peer feedback and its impact on learning.

Q1. The student regularly prepared for tutorials
Q2. The student participated actively in tutorials
Q3. The student showed behavior and input that facilitated my learning
Q4. The student was punctual to PBL tutorials
Q5. The student listened to and showed respect for the opinions of others
Q6. The student brought in new information to share with the group
Q7. The student was able to communicate ideas clearly
Q8. The student gave input that was focused and relevant to the case
Q9. The student accepted and responded to criticism gracefully
Q10. During this block, working with my PBL group facilitated my learning

Figure 1 Modified scale for the peer assessment of professional learning behaviour in a PBL group (checklist items; 1-9 and global rating; item 10)

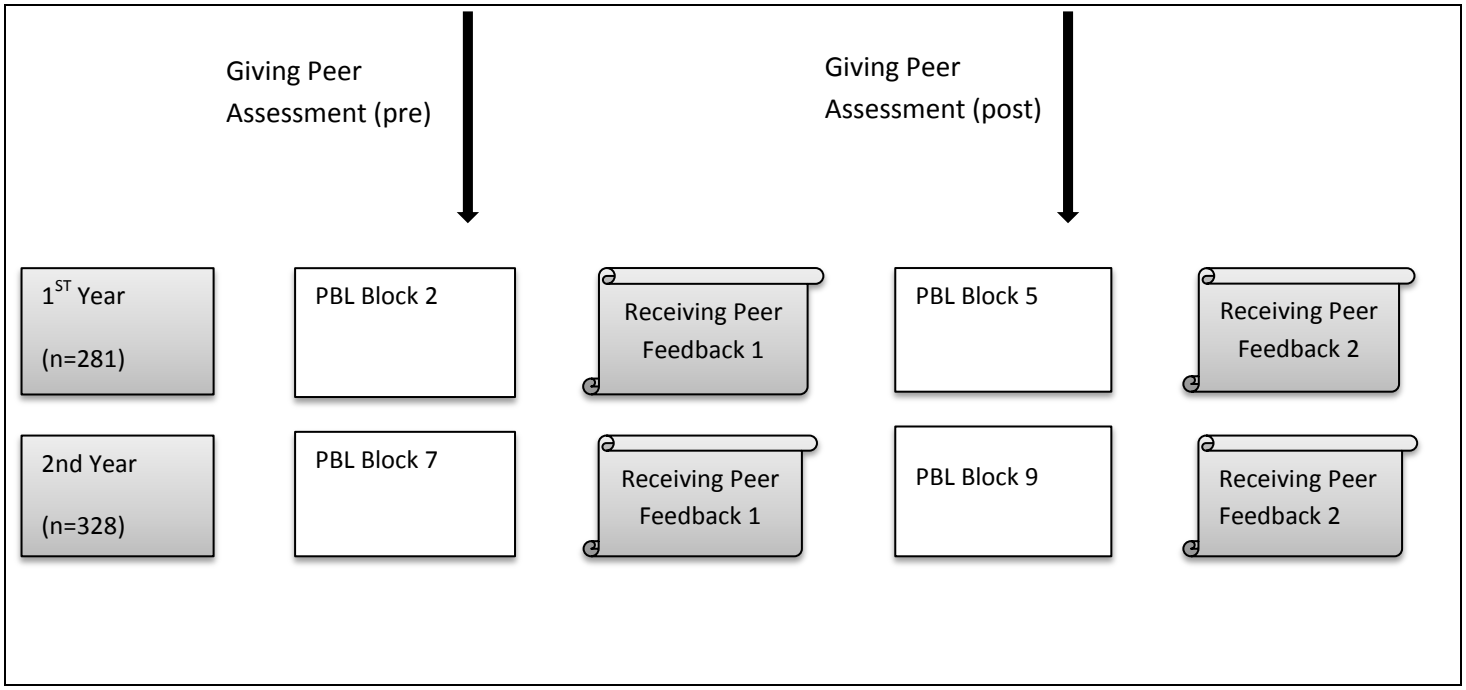

Figure 2 Quasi experimental pre- post- test design for two cohorts in academic years one (n= 305) and two (n=328) during a single calendar year of giving PBL peer assessment with the intervention being the receiving of summarized peer feedback.

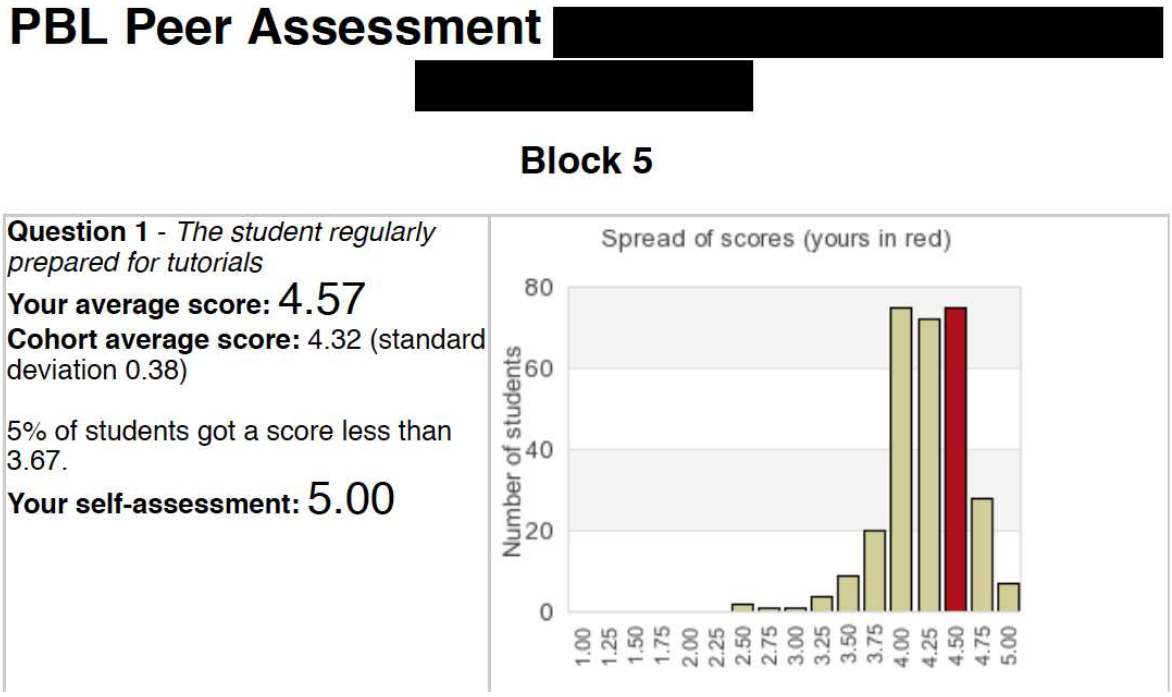

Figure 3 Anonymised sample of individualised feedback on the first checklist item from the nine-item scale using cohort averaged scores

		Pre- (Occasion one)		Post- (Occasion two-)	
Component	**Meaning**	**Estimate**	**Proportion**	**Estimate**	**Proportion**
Var_p	Student professional behaviour	.047	16%	.062	19%
Var_r	Assessor stringency/leniency	.163	56%	.168	51%
Var_g	PBL Group	.016	5%	.022	6%
Var_{pj}	Assessor subjectivity	.065	22%	.081	24%

Table 1 Variance components of the peer assessment of professional behaviours of 1st Year students (n=305)

		Pre- (occasion one)		Post- (occasion two)	
Component	**Meaning**	**Estimate**	**Proportion**	**Estimate**	**Proportion**
Var_p	Student professional behaviour	.028	11%	.035	13%
Var_r	Assessor stringency/leniency	.167	64%	.159	59%
Var_g	PBL Group	.007	3%	.009	3%
Var_{pj}	Assessor subjectivity	.058	22%	.065	24%

Table 2 Variance components of the peer assessment of professional of 2nd year students (n=328)

Dependability

	Year 1				**Year 2**			
	Across groups		**Within Groups**		**Across groups**		**Within Groups**	
	Pre-	**Post-**	**Pre-**	**Post-**	**Pre-**	**Post-**	**Pre-**	**Post-**
G coefficient (G)	**0.53**	**0.56**	**0.87**	**0.87**	**0.47**	**0.51**	**0.81**	**0.83**
Standard Error of Measurement (SEM)	**0.74**	**0.75**	**0.31**	**0.32**	**0.69**	**0.69**	**0.31**	**0.32**

Table 3 D study modeling changes in reliability for groups of ten students when considering their professional behaviour scores across groups and within groups both before and after they received feedback on their own PBL performance for 1st and 2nd years.

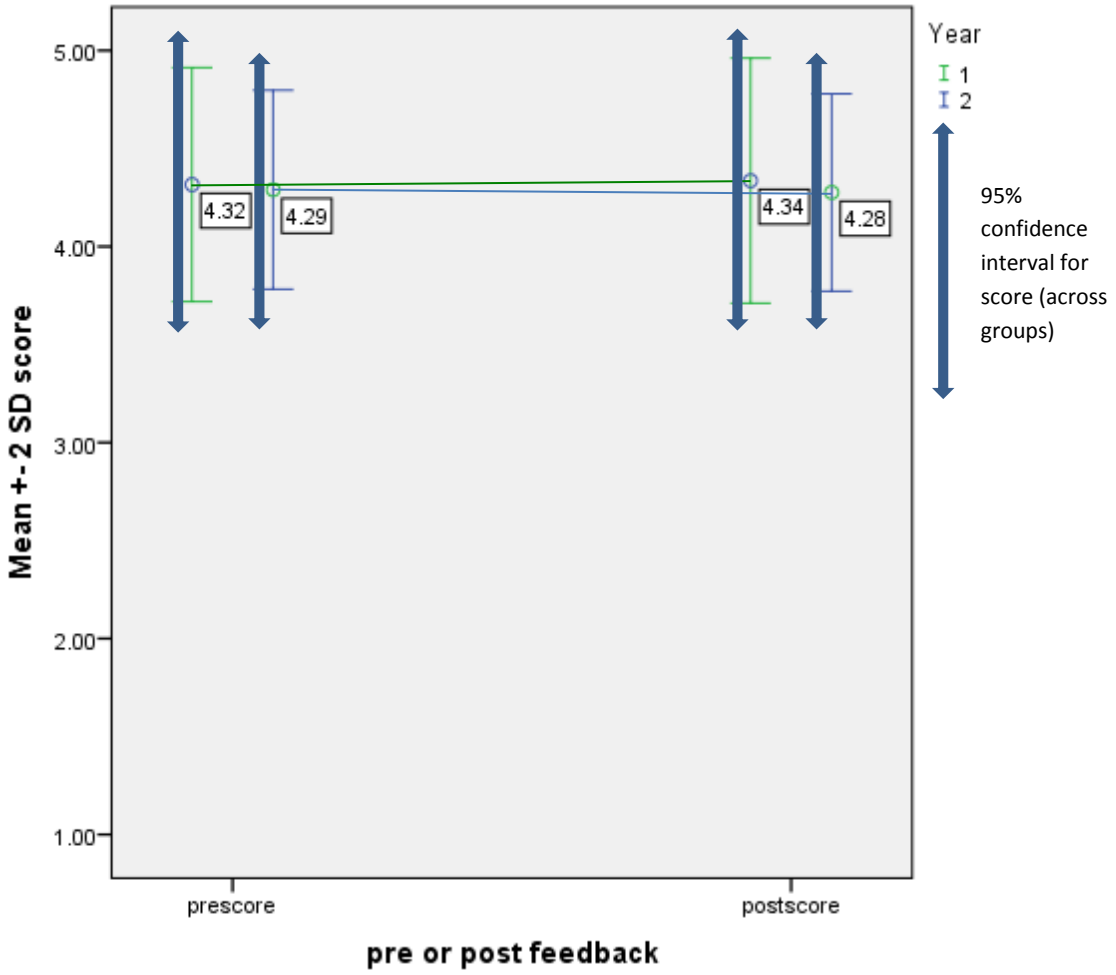

Figure 4 A combined figure showing mean scores on the professional learning behaviours scale for 1st and 2nd years, both before and after receiving standardised peer feedback. Standard errors of measurement (SEM) have been placed around the mean scores, as well as standard deviations (SD).

References

[1] DesRoches C, Rao S, Fromson J, et al. Physicians' Perceptions, Preparedness for Reporting, and Experiences Related to Impaired and Incompetent Colleagues. *JAMA*. 2010;**304**:187-193.

[2] Rosenthal M. *The Incompetent Doctor. Behind Closed Doors.* Buckingham Open University Press; 1995.

[3] Wright C, Richards SH, Hill JJ, et al. Multisource Feedback in Evaluating the Performance of Doctors: The Example of the UK General Medical Council Patient and Colleague Questionnaires. *Acad Med*. 2012;**87**:1668-1678.

[4] Australian Medical Council Working Party. Good Medical Practice: A code of conduct for doctors in Australia. Australian Medical Council Ltd, Canberra, 2009.

[5] Abdulla A. A critical analysis of mini peer assessment tool (mini- PAT) *JR Soc Med* 2008;**101**:22-26.

[6] Hawkins R, Katsufrakis P, Holtman M, Clauser B. Assessment of Medical Professionalism: Who, what, when, where, how, and ...why? *Med Teach*. 2009;**31**:348-361.

[7] Lockyer J. Multisource (360 degree) feedback and the assessment of ACGME competencies for Emergency Medicine Residents. Faculty of Medicine, University of Calgary, Calgary, 2010.

[8] Zhao Y, Zhang X, Chang Q, Sun B. Psychometric Characteristics of the 360° Feedback Scales in Professionalism and Interpersonal and Communication Skills Assessment of Surgery Residents in China. *J Surg Educ*. 2013;**70**:628-635.

[9] Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts KM. Evaluation of physicians' professional performance: an iterative development and validation study of multisource feedback instruments. *BMC Health Serv Res*. 2012;**12**:80.

[10] Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ*. 2010;**341**.

[11] Papadakis MA, Hodgson CS, Teherani A, Kohatsu ND. Unprofessional Behavior in Medical School Is Associated with Subsequent Disciplinary Action by a State Medical Board. *Acad Med*. 2004;**79**:244-249.

[12] Papadakis MA, Loeser H, Healy K. Early Detection and Evaluation of Professionalism Deficiencies in Medical Students: One School's Approach. *Acad Med*. 2001;**76**:1100-1106.

[13] Arnold L, Shue C, Kalishman S, et al. Can there be a single system for peer assessment of professionalism among medical students? A multi-institutional study. *Acad Med*. 2007;**82**:578-586.

[14] Arnold L, Shue CK, Kritt B, Ginsburg S, Stern DT. Medical Students' Views on Peer Assessment of Professionalism. *J Gen Intern Med*. 2005;**20**:819-824.

[15] Sandars J. the use of reflection in medical education: AMEE Guide No. 44. *Med Teach*. 2009;**31**:685-695.

[16] Boud D, Molloy E. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education*. 2013;**38**:698-712.

[17] Kamp RA, Dolmans DJM, Van Berkel HM, Schmidt H. The effect of midterm peer feedback on student functioning in problem-based tutorials. *Advances in Health Sciences Education*. 2013;**18**:199-213.

[18] Rees C, Shepherd M. The acceptability of 360-degree judgements as a method of assessing undergraduate medical students' personal and professional behaviours. *Med Educ*. 2005;**39**:49-57.

[19] Lee K-L, Tsai S-L, Chiu Y-T, Ho M-J. Can student self-ratings be compared with peer ratings? A study of measurement invariance of multisource feedback. *Advances in Health Sciences Education*. 2015:1-13.

[20] Eva KW. Assessing tutorial-based assessment. *Advances in Health Sciences Education*. 2001;**6**:243-257.

[21] Papinczak T, Young L, Groves M, Haynes M. An analysis of peer, self, and tutor assessment in problem-based learning tutorials. *Medical Teacher*. 2007;**29**:e122-e132.

[22] Kamp RJA, Dolmans DHJM, Van Berkel HJM, Schmidt HG. Can students adequately evaluate the activities of their peers in PBL? *Med Teach*. 2011;**33**:145-150.
[23] Van Mook WNKA, De Grave WS, Huijssen-Huisman E, et al. Factors inhibiting assessment of students' professional behaviour in the tutorial group during problem-based learning. *Med Educ*. 2007;**41**:849-856.
[24] Shue C, Arnold L, Stern D. Maximizing Participation in Peer Assessment of Professionalism: The Students Speak *Acad Med*. 2005;**80**:S1-S5.
[25] Nofziger A, Naumburg E, Davis B, Mooney C, Epstein R. Impact of Peer Assessment on the Professional Development of Medical Students: A Qualitative Study. *Academic Medicine*. 2010;**85**:140-147.
[26] Visschers-Pleijers ASF, Dolmans DJM, Wolfhagen IAP, Vleuten CMV. Student Perspectives on Learning-Oriented Interactions in the Tutorial Group. *Advances in Health Sciences Education*. 2005;**10**:23-35.
[27] S. de Grave DHJMD, Cees P.M. van der Vleuten, Willem. Student perceptions about the occurrence of critical incidents in tutorial groups. *Med Teach*. 2001;**23**:49-54.
[28] Khoiriyah U, Roberts C, Jorm C, Van der Vleuten C. Enhancing students' learning in problem based learning: validation of a self-assessment scale for active learning and critical thinking. *BMC Med Educ*. 2015;**15**:140.
[29] Denchfield-Schlecht K. Feedback. *Int Anesthesiol Clin*. 2008;**46**:67-84.
[30] Kamp RA, van Berkel HM, Popeijus H, Leppink J, Schmidt H, Dolmans DJM. Midterm peer feedback in problem-based learning groups: the effect on individual contributions and achievement. *Advances in Health Sciences Education*. 2014;**19**:53-69.
[31] Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;**49**:560-575.
[32] van Mook WKA, Muijtjens AM, Gorter S, Zwaveling J, Schuwirth L, van der Vleuten CM. Web-assisted assessment of professional behaviour in problem-based learning: more feedback, yet no qualitative improvement? *Advances in Health Sciences Education*. 2012;**17**:81-93.
[33] Papinczak T, Young L, Groves M, Haynes M. An analysis of peer, self, and tutor assessment in problem-based learning tutorials. *Med Teach*. 2007;**29**:e122-e132.
[34] Reiter HI, Eva KW, Hatala RM, Norman GR. Self and peer assessment in tutorials: application of a relative-ranking model. *Acad Med*. 2002;**77**:1134-1139.
[35] Sullivan ME, Hitchcock MA, Dunnington GL. Peer and self assessment during problem-based tutorials. *Am J Surg*. 1999;**177**:266-269.
[36] Crossley J, Russell J, Jolly B, et al. 'I'm pickin' up good regressions': the governance of generalisability analyses. *Med Educ*. 2007;**41**:926-934.
[37] Roberts C, Rothnie I, Zoanetti N, Crossley J. Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Med Educ*. 2010;**44**:690-698.
[38] Field MJ, Sefton AJ. Computer-based management of content in planning a problem-based medical curriculum. *Med Educ*. 1998;**32**:163-171.
[39] Hendry GD, Lyon PM, Prosser M, Sze D. Conceptions of problem-based learning: the perspectives of students entering a problem-based medical program. *Med Teach*. 2006;**28**:573-575.
[40] Langendyk V. Not knowing that they do not know: self-assessment accuracy of third-year medical students. *Med Educ*. 2006;**40**:173-179.
[41] Epstein R. Learning from the problems of problem-based learning. *BMC Med Educ*. 2004;**4**:1.
[42] Moust JHC, Berkel HJMv, Schmidt HG. Signs of Erosion: Reflections on Three Decades of Problem-Based Learning at Maastricht University. *Higher Education*. 2005;**50**:665-683.
[43] Cronbach LJ, Glaser GC, Nanda H, Rajaratnam N. *The dependability of behavioural measurements: the theory of generalisability for scores and profiles*. New York John Wiley; 1972.
[44] Brennan RL. The Conventional Wisdom About Group Mean Scores. *Journal of Educational Measurement*. 1995;**32**:385-396.

[45] Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Med Educ*. 2011;**45**:560-569.
[46] Roberts C, Walton M, Rothnie I, et al. Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Med Educ*. 2008;**42**:396-404.
[47] Zhang B, Johnston L, KIlic G. Assessing the relaibility of self- and peer rating in student group work. *Asessment & Evaluation in Higher Education*. 2008;**33**:329-340.
[48] van Zundert M, Sluijsmans D, van Merriënboer J. Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*. 2010;**20**:270-279.
[49] Sluijsmans DMA, Brand-Gruwel S, van Merriënboer JJG. Peer Assessment Training in Teacher Education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education*. 2002;**27**:443-454.
[50] Crossley J, Jolly B. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Med Educ*. 2012;**46**:28-37.
[51] Dolmans DHJM, Wolfhagen IHAP, Van Der Vleuten CPM, Wijnen WHFW. Solving problems with group work in problem-based learning: hold on to the philosophy. *Med Educ*. 2001;**35**:884-889.
[52] Hendry GD. Tutors perception of dysfunctional behaviour in problem based learning tutorial groups. *HERSDA News*. 2002;**24**:27-30.
[53] Dolmans DHJM, De Grave W, Wolfhagen IHAP, Van Der Vleuten CPM. Problem-based learning: future challenges for educational practice and research. *Med Educ*. 2005;**39**:732-741.
[54] De Grave WS, Dolmans DH, Van Der Vleuten CP. Student perspectives on critical incidents in the tutorial group. *Adv Health Sci Educ Theory Pract*. 2002;**7**:201-209.

Author Contributions

CJ conceived of and with the assistance of SG implemented the Peer Assessment Process within PBL. SG extracted the data, CR and JC developed the research design, and JC undertook analysis, and led the interpretation. CR wrote the first draft of the manuscript, and all contributed to revisions, and approved the final manuscript.