Tutor assessment of medical students in problem-based learning sessions

Bader Khawaji[1,2], Emad Masuadi[3], Abdulrahman Alraddadi[4,5],
Muhammad Anwar Khan[2,6], Syed Sameer Aga[1,2], Hatim Al-Jifree[2,7],
Mohi Eldin Magzoub[3]

Abstract:

BACKGROUND: Problem-based learning (PBL) is a method of learning that has been adopted in different curricula of different disciplines for more than 30 years; the assessment of the students in PBL sessions in medical schools is fundamental to ensure students' attainment of the expected outcomes of conducting PBL sessions and in providing the students with the feedback that help them to develop and encourage their learning. This study investigated the inter-rater reliability of the tutor assessment in assessing medical students' performance in their PBL tutorial sessions.

MATERIALS AND METHODS: This study was conducted in the College of Medicine (COM), in the academic year 2021–2022. The study involved ten raters (tutors) of two genders who assessed 33 students in three separate PBL tutorial sessions. The PBL sessions were prerecorded and shown to the 10 raters for their assessment of PBL sessions.

RESULTS: This study showed that male raters gave higher scores to students compared with female raters. In addition, this investigation showed low inter-rater reliability and poor agreement among the raters in assessing students' performance in PBL tutorial sessions.

CONCLUSION: This study suggests that PBL tutor assessment should be reviewed and evaluated; this should be performed with consideration of using assessment domains and criteria of performance. Thus, we recommend that 360-degree assessment including tutor, self, and peer assessment should be used to provide effective feedback to students in PBL tutorial sessions.

Keywords:
Assessment, medical curriculum, problem-based learning (PBL)

[1]Department of Basic Medical Sciences, College of Medicine, King Saud Bin Abdulaziz University for Health Sciences (KSAU-HS), Jeddah, Saudi Arabia, [2]King Abdullah International Medical Research Center (KAIMRC), Ministry of National Guard, Health Affairs, Jeddah, Saudi Arabia, [3]College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, United Arab Emirates, [4]King Abdullah International Medical Research Center (KAIMRC), Ministry of National Guard, Health Affairs, Riyadh, Saudi Arabia, [5]Department of Basic Medical Sciences, College of Medicine, King Saud Bin Abdulaziz University for Health Sciences (KSAU-HS), Riyadh, Saudi Arabia, [6]Department of Medical Education, College of Medicine, King Saud Bin Abdulaziz University for Health Sciences, Jeddah, Saudi Arabia, [7]Department of Oncology, King Abdulaziz Medical City, Ministry of National Guard Health Affairs, Jeddah, Saudi Arabia

Address for correspondence:
Dr. Bader Khawaji, Department of Basic Medical Sciences, College of Medicine, King Saud Bin Abdulaziz University for Health Sciences P.O. BOX 9515, Jeddah 21423, MC 6677, Saudi Arabia. E-mail: khawajib@ksau-hs. edu.sa

Received: 05-09-2023
Accepted: 04-12-2023
Published: 11-07-2024

Introduction

Problem-based learning (PBL) is an interactive method of student learning that has been adopted in different curricula, which is based on problems, that are required to be solved through research, self-learning, and knowledge application.[1,2] PBL teaches students effective learning skills, teamwork, critical thinking, and problem-solving, in addition to knowledge application and retention—compared with the traditional instructional method of learning.[3,4] Therefore, many medical schools around the globe have adopted the PBL method

in their instructive methods,[5,6] and so did ours—the College of Medicine (COM) at King Saud bin Abdulaziz University for Health Sciences (KSAU-HS) in Jeddah, Saudi Arabia.

The PBL assessment provides information about competencies and skills that are not usually assessed by traditional assessments[7], as well as being important in encouraging the students' learning.[8] Moreover, the assessment in the PBL tutorial sessions plays an important role in providing feedback to the students, helping them to develop their skills in decision-making and critical thinking.[5] However, the assessment of the students in PBL sessions by the tutors

can be challenging, due to the difficulty in ensuring that individual students have acquired the desired outcomes and competencies.[7] There are three types of assessment of the students' performance in PBL tutorial sessions; these are self-assessment,[5,7,9,10] peer assessment,[5,9,11,12] and tutor assessment.[5,9]

For any PBL-based curriculum, an assessment of the PBL sessions is still one of the key challenges affecting the overall grading, as different institutions are using different assessment tools ranging between scale grading systems and criteria of the performance of the students.[9] It has long been argued that the rating-based assessment may be affected by the limitation of the human judgment perception limitation, mental workload, and the cognitive efforts of the rater.[13] Moreover, the global rating assessment of PBL may be less reliable compared with using domain-based forms of assessment. Hence, no more than four domains should be used; this is due to the difficulty of assessing the students' performance using too many criteria or domains by the tutors.[6]

The correlation between self-assessment, peer assessment, and tutor assessment has been found that a student's self-assessment is usually lower as compared to that of the tutor, while peers tend to give a better rating to their colleagues compared with both self-assessment or that of a PBL tutor.[14] Peer assessment has been argued to provide appropriate feedback and assessment of the students' performance in the PBL tutorial sessions; however, students were in more agreement about their peers' assessment.[12] In addition, the self-assessment and peer assessment are still questionable and may not be valid to be included in a summative form of assessment.[10] Moreover, it has been reported that tutors may believe that students are undeveloped to be able to assess their own performance in the PBL tutorial sessions.[9]

One of the assessment principles in conducting problem-based tutorial sessions is the formative function of assessment that has not been examined concerning the outcomes that result from this type of learning method. This was reported in conjunction with the diversity and uniformity of PBL assessment with the varied curricula and circumstances within the different educational institutions that adopt the PBL method.[15] For instance, one review showed that tutors may be undermining students' performance and being subjective when they assess students in PBL tutorials using an assessment method that was mainly descriptive to target giving feedback to students.[16] In addition, the assessment tools used by tutors have not been investigated previously about the correlation and reliability of the tools produced regardless of the tutors, although our group did report on the quality of the PBLs implemented in the curriculum.[17]

As a faculty who is facilitating PBL tutorial sessions in the COM at KSAU-HS in Jeddah, it has been observed that some students reported that there are some inconsistencies in the produced scores using the current assessment tool. It is fundamental to define the reliability and accuracy of the tools for assessing students' performance in PBL tutorial sessions. This study aimed to examine and assess the inter-rater reliability among different tutors assessing students' performance in PBL tutorial sessions.

Materials and Methods

Study design and setting
The study was conducted at the COM, KSAU-HS, in Jeddah, Saudi Arabia, in the academic year 2021–2022. The current research is an analytical cross-sectional study that examines the inter-rater reliability among faculty members assessing the students in PBL tutorial sessions.

Study participants and sampling
The curriculum in the COM is PBL based, as students are involved in PBL tutorial sessions as a method of learning and teaching. Faculty members are involved in the PBL sessions as tutors assess students.

Data collection tool and technique
The instrument that was used for the data collection was the PBL assessment tool (sheet) that has been used to assess students' performance in PBL tutorial sessions at the COM, KSAU-HS, Jeddah, Saudi Arabia, and is available on request. It consists of 1 to 5 grading, which was configured as 1: poor, 2: below average, 3: average, 4: good, and 5: outstanding. The assessment tool provided to each rater was in the English language.

The study involved ten tutors, five males, and five females, who served as raters for the prerecorded PBL sessions. For the actual PBL session, 33 students participated in three separate groups, and each group was provided with the medical case. The sessions of all three groups were recorded after consent for participation was obtained from the students. The raters assessed the students' performance based on a global scale rating. This study ensured several ethical considerations.

The participants in this study (PBL tutors) were invited to a lecture hall to fill out the assessment tool. There were three videos recorded in PBL sessions that were viewed by the participants; each video was 25 minutes long showing a real scenario of a session of PBL tutorial, which involved comprehensive discussion and preparation by students. Participants were given an assessment tool to fill with each PBL session to assess the students' performance; then, they were given 10 minutes after viewing each video to ensure the completion of the assessment tool (available on request).

The collected data were analyzed using IBM Statistical Package for the Social Sciences (SPSS) Statistics software (version 21 © Copyright IBM Corporation and other(s) 1989, 2012). The mean and standard deviation (SD) were calculated for the assessment scores, while a one-way analysis of variance (ANOVA) was used to assess differences in the mean scores that were given by each rater for all three PBL groups. Additionally, the paired-samples *t*-test was utilized to compare quantitative variables such as those between students' scores between male and female raters. In addition, interclass correlation (ICC) and Krippendorff's alpha and weighted kappa test were used to assess the reliability and degree of agreement of the ratings among the PBL raters. Finally, all statistical tests were considered significant when $P < 0.05$.

Ethical consideration

The study was conducted after due ethical approval and IRB from King Abdullah International Medical Research Center (KAIMRC: No.: SP18472J). Additionally, before video recording of the PBL sessions, consent from students to record their PBL sessions and have them used for research purposes was taken. Finally, the consent of the faculty members to participate in this study was obtained before their participation.

Results

Table 1 shows the scores that were given to the 33 students by the 10 raters. This shows the different scores that were given to the same students by the different raters (five males and five females). The average score of the rating was 3.75 (range: 2.10 to 4.90; SD: 0.73); male raters gave a mean score of 4.07 (range: 2.40 to 5.0; SD: 0.64), while female raters gave an average score of 3.44 (range: 1.80 to 5.0; SD: 0.84).

The differences in the mean scores that were given by each rater among the three PBL groups are shown in Table 2. It has been seen from the results of this study that one-way ANOVA indicated that there were no significant differences in the mean ratings by each rater among the three different PBL groups. These results showed that students in PBL group 1 scored on average 3.7 ± 0.73, while the second group had a rating of 3.94 ± 0.73 and the third group scored 3.63 ± 0.69. These results ($P = 0.557$) suggest no significant differences in the average given scores to the students in each PBL group; this was reported as well to each of the 10 raters.

Differences in the given scores between male and female raters are shown in Figure 1. Figure 1 shows that male raters gave higher scores (4.07 ± 0.64) to students compared with the female raters, who gave an average score of 3.44 ± 0.84 to the students ($P < 0.001$).

Table 1: Students' scores that were given by the different raters (R: rater)

Students	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Overall
1	4	5	5	5	5	4	4	3	4	3	4.20
2	4	5	5	4	5	5	5	3	3	4	4.30
3	4	5	5	4	5	5	4	2	3	4	4.10
4	3	1	4	4	4	5	3	1	3	3	3.10
5	2	1	4	3	2	2	1	1	3	2	2.10
6	3	2	5	4	5	5	5	1	4	4	3.80
7	3	2	4	5	5	4	3	4	3	3	3.60
8	5	2	5	4	4	3	2	3	3	5	3.60
9	5	5	5	4	5	5	5	5	5	5	4.90
10	4	3	4	4	4	4	3	1	3	4	3.40
11	4	5	5	3	5	3	2	1	3	5	3.60
12	4	5	5	5	5	4	5	5	4	4	4.60
13	5	5	5	5	5	5	5	4	5	5	4.90
14	3	3	4	4	5	2	2	3	3	3	3.20
15	4	3	5	3	3	3	3	3	2	3	3.20
16	4	3	5	4	5	4	4	3	4	4	4.00
17	5	5	5	5	5	5	4	4	5	4	4.70
18	5	5	5	3	5	5	5	5	4	5	4.70
19	4	2	4	4	4	4	3	3	2	2	3.20
20	5	4	5	4	4	4	4	4	4	4	4.20
21	3	2	5	4	2	3	3	2	2	3	2.90
22	4	5	5	4	5	3	3	1	2	5	3.70
23	3	5	5	5	5	3	5	3	3	4	4.10
24	4	3	4	4	5	4	4	3	2	5	3.80
25	4	4	5	3	5	3	4	4	3	5	4.00
26	2	1	4	4	4	2	2	2	2	2	2.50
27	3	2	4	4	3	2	2	1	2	2	2.50
28	3	5	4	5	5	4	5	4	3	4	4.20
29	4	5	5	4	5	4	5	3	2	5	4.20
30	4	1	4	4	4	3	2	2	2	3	2.90
31	3	4	5	4	5	3	4	4	4	4	4.00
32	4	4	5	3	5	5	5	5	5	2	4.30
33	4	3	5	3	3	2	3	4	4	3	3.40

Table 2: Differences in the mean scores that were given by each rater among the three PBL groups

	PBL group						*P*
	1		**2**		**3**		
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Rater 1	3.73	0.90	4.18	0.75	3.45	0.69	0.108
Rater 2	3.27	1.74	3.82	1.25	3.36	1.50	0.666
Rater 3	4.64	0.50	4.82	0.40	4.55	0.52	0.409
Rater 4	4.00	0.63	4.09	0.70	3.91	0.70	0.822
Rater 5	4.45	0.93	4.36	1.03	4.45	0.82	0.966
Rater 6	4.09	1.04	3.82	0.98	3.18	0.98	0.11
Rater 7	3.36	1.36	3.73	1.01	3.73	1.27	0.726
Rater 8	2.27	1.42	3.36	1.21	3.18	1.17	0.115
Rater 9	3.36	0.67	3.36	1.21	2.91	1.04	0.478
Rater 10	3.82	0.98	3.82	0.98	3.55	1.21	0.788
Overall	3.7	0.73	3.94	0.73	3.63	0.69	0.577

Moreover, the inter-rater reliability among the raters according to their rating was examined using the ICC and Krippendorff's alpha statistical tests. From the
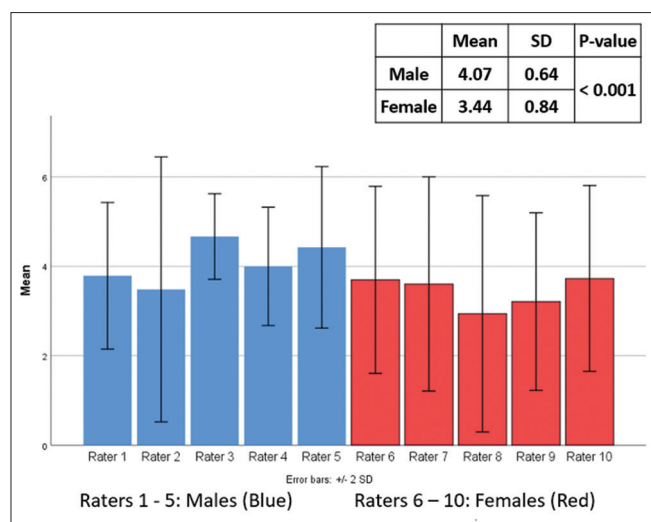
	Mean	SD	P-value
Male	4.07	0.64	< 0.001
Female	3.44	0.84	

Raters 1 - 5: Males (Blue)   Raters 6 – 10: Females (Red)

Figure 1: Differences in the given scores between male and female raters

Table 3: Inter-rater reliability among the raters using the interclass correlation (ICC) and Krippendorff's alpha

	Intraclass correlation (ICC)	P	Krippendorff's alpha
Overall	0.40	<0.001	0.31
Male raters	0.34	<0.001	0.25
Female raters	0.45	<0.001	0.42

Table 4: Level of agreement in the ratings between the different raters

	Kappa agreement *P<0.05									
Rater	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.25*	1.00								
3	0.12	0.14*	1.00							
4	-0.05	0.13	0.03	1.00						
5	0.25*	0.55*	0.11	0.15	1.00					
6	0.38*	0.26*	0.11	0.13	0.53*	1.00				
7	0.13	0.36*	0.27*	0.29*	0.32*	0.30*	1.00			
8	0.16*	0.37*	0.05	-0.06	0.39*	0.22*	0.12*	1.00		
9	0.26*	0.27*	0.05	0.04	0.35*	0.34*	0.11	0.38*	1.00	
10	0.40*	0.41*	0.11	0.02	0.41*	0.28*	0.31*	0.17	0.22*	1.00

results that are presented in Table 3, it can be seen that ICC (r = 0.40, *P* < 0.001) and Krippendorff's alpha (0.31) show poor reliability among the raters in assessing the students' performance in the video-recorded PBL tutorial sessions.

Finally, the weighted kappa agreement statistical test [Table 4] was used to measure the level of agreement in rating between the different raters. The table shows a number of slightly fair agreements between some of the raters. This can be seen, for example, between rater 5 with raters 2 and 6 and rater 10 with raters 2 and 5. In addition, most of the raters had poor agreement with each other. Therefore, the kappa agreement test suggests poor-to-moderate agreement between the different raters in assessing the students' performance in the PBL tutorial sessions.

Discussion

In reviewing the literature, it has been reported that students' learning is primarily affected by the assessment[18]; this suggests that students' performance in the problem-based tutorial sessions may be driven and triggered by the method of the assessment used. Therefore, an objective assessment of students in PBL tutorial sessions has been emphasized to achieve the objectives of conducting PBL tutorial sessions.[5] Tutor assessment is an important method that assesses the performance of students in PBL tutorial sessions, which will serve as good feedback to the students about their performance and areas for improvement.[5,9] In addition, the tutor assessment type of PBL assessment has been reported to be challenging and not sufficient, especially in ensuring students' acquisition of the required objectives and skills.[7] However, in the literature, no data on the inter-rater reliability among raters who assess students' performance in PBL tutorial sessions have been reported.

The results of the current study suggest that there is poor inter-rater reliability among the raters in assessing the students' performance in the PBL tutorial sessions. However, the mean average scores that were given to each PBL group were found to have no differences among the different raters; this might be explained by the fact that students at the COM at KSAU-HS are allocated and distributed to their PBL groups evenly according to their grade point averages (GPAs).

Surprisingly, the five female raters were found to give lower scores to students than the five male raters. One possible explanation for these results is that female raters usually do not assess male students and they mainly assess female students at KSAU-HS, and they may compare the male students to their female peers. The current study compared the scores of the PBL assessment that were given by the 10 raters to all the 33 students; these scores that were given to each student were analyzed using three different statistical tests; these are ICC, Krippendorff's alpha, and the kappa agreement test. These statistical tests indicated poor inter-rater reliability and poor agreement among the 10 raters in assessing the 33 students in their PBL tutorial sessions.

The results of the current study agree with what has been suggested in the literature; this includes the theories of the subjectivity of the tutor assessment of the PBL tutorial session (16). This has been suggested, while diversity and uniformity of rating-based assessments of the PBL have been used around the world.[15] Human judgment perception limitation, mental workload, and the cognitive efforts of the raters are among the suggested reasons for

this variability.[13] Therefore, it has been suggested that using a domain-based assessment type may be more reliable than a global rating-based assessment; however, it has been suggested that there should not be too many domains or criteria of performance that may affect the ability of the rater to assess the students according to the relevant domains.[6]

Tutor assessment showed several drawbacks; therefore, it should be used with criteria of performance to reduce subjectivity. In addition, self-assessment[5,7,9,10] and peer assessment[5,9,12] are useful methods to give feedback as well, as long as no summative value should be utilized.[10]

Limitations

The current study may have a number of limitations; these include the sample size that has been used in investigating the inter-rater reliability among the raters assessing the students in PBL tutorial sessions. In addition, the three PBL groups that were video recorded were conducting the first session of a PBL scenario; this may affect the ability of the tutor to assess some of the student's components of their performance. Moreover, all of the students who were video recorded in the PBL sessions that were conducted were male students; in future investigations, female students should be included. Finally, the current study was conducted in a single institution, and a better result in future investigations might be obtained with collaboration with other institutions, as this may provide more consistent and detailed data on the tutor assessment of the students in PBL tutorial sessions.

Recommendation

This study advises that the assessment domains and performance criteria be considered for use in tutor-based PBL evaluation of students' sessions. Furthermore, it is proposed that tutor, self, and peer evaluation approaches be used in tandem to meet the goals of providing effective feedback to students in PBL tutorial sessions.

Conclusion

The tutor assessment of the students in PBL sessions is an important method to give feedback to students to develop and encourage their learning. In the current study, the aim was to assess the inter-rater reliability of assessing the students in PBL tutorial sessions. This study has shown that there is low inter-rater reliability and poor agreement among the raters in assessing students' performance in PBL tutorial sessions. The current study highlights the importance of reviewing and evaluating the current PBL tutor assessment and the rating-based system that has been used. Finally, a further extended investigation should be conducted to establish the validity and inter- and intra-rater reliability of the PBL tutor assessment method; this investigation should

have a larger sample size from both genders, assessing different assessment tools (forms) and domains (criteria) of performance.

Ethical consideration

The Institutional Review Board of the King Abdullah International Medical Research Center approved this study (KAIMRC: No.: SP18472J).

Conflicts of interest

There are no conflicts of interest.

References

1. Walker AE, Leary H, Hmelo-Silver CE, Ertmer PA. Essential Readings in Problem-Based Learning. West Lafayette: Purdue University Press; 2015.
2. Duch BJ, Groh SE, Allen DE. The Power of Problem-Based Learning: A Practical "How to" for Teaching Undergraduate Courses in any Discipline. Sterling: Stylus Publishing LLC; 2001.
3. Alrahlah A. How effective the problem-based learning (PBL) in dental education. A critical review. Saudi Dent J 2016;28:155-61.
4. Yew EH, Goh K. Problem-based learning: An overview of its process and impact on learning. Health Prof Educ 2016;2:75-9.
5. Elizondo-Montemayor LL. Formative and summative assessment of the problem-based learning tutorial session using a criterion-referenced system. J Int Assoc Med Sci Educ 2004;14:8-14.
6. Lee M, Wimmers PF. Validation of a performance assessment instrument in problem-based learning tutorials using two cohorts of medical students. Adv Health Sci Educ 2016;21:341-57.
7. Eva KW. Assessing tutorial-based assessment. Adv Health Sci Educ 2001;6:243-57.
8. Wood DF. ABC of learning and teaching in medicine: Problem based learning. BMJ 2003;326:328.
9. Bollela VR, Gabarra MH, da Costa C, Lima RC. Students and tutors' social representations of assessment in problem-based learning tutorials supporting change. BMC Med Educ 2009;9:30.
10. Machado JL, Machado VM, Grec W, Bollela VR, Vieira JE. Self-and peer assessment may not be an accurate measure of PBL tutorial process. BMC Med Educ 2008;8:55.
11. Papinczak T, Young L, Groves M. Peer assessment in problem-based learning: A qualitative study. Adv Health Sci Educ 2007;12:169-86.
12. Kritikos VS, Woulfe J, Sukkar MB, Saini B. Intergroup peer assessment in problem-based learning tutorials for undergraduate pharmacy students. Am J Pharm Educ 2011;75:73.

13. Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. Adv Health Sci Educ 2013;18:291-303.

14. Papinczak T, Young L, Groves M, Haynes M. An analysis of peer, self, and tutor assessment in problem-based learning tutorials. Med Teach 2007;29:122-32.

15. Nendaz MR, Tekian A. Assessment in problem-based learning medical schools: A literature review. Teaching and Learning in Medicine 1999;11:232-43.

16. Allareddy V, Havens AM, Howell TH, Karimbux NY. Evaluation of a new assessment tool in problem-based learning tutorials in dental education. J Dent Educ 2011;75:665-71.

17. Abed SS, Al-Mansour M, Ahmed SN, Khan MA, Martin PN, Shah SW, *et al*. Evaluation of problem-based learning in college of medicine: An indicator of quality education in a hybrid curriculum. Educ Res Int 2023;2023. doi: 10.1155/2023/9955409.

18. Ghaicha A. Theoretical framework for educational assessment: A synoptic review. J Educ Pract 2016;7:212-31.