ORIGINAL RESEARCH

Validity and reliability assessment of a peer evaluation method in team-based learning classes

Hyun Bae Yoon[1], Wan Beom Park[1,2], Sun-Jung Myung[1], Sang Hui Moon[1] and Jun-Bean Park[1]

[1]Medical Education Office and [2]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Korea

Purpose: Team-based learning (TBL) is increasingly employed in medical education because of its potential to promote active group learning. In TBL, learners are usually asked to assess the contributions of peers within their group to ensure accountability. The purpose of this study is to assess the validity and reliability of a peer evaluation instrument that was used in TBL classes in a single medical school.

Methods: A total of 141 students were divided into 18 groups in 11 TBL classes. The students were asked to evaluate their peers in the group based on evaluation criteria that were provided to them. We analyzed the comments that were written for the highest and lowest achievers to assess the validity of the peer evaluation instrument. The reliability of the instrument was assessed by examining the agreement among peer ratings within each group of students via intraclass correlation coefficient (ICC) analysis.

Results: Most of the students provided reasonable and understandable comments for the high and low achievers within their group, and most of those comments were compatible with the evaluation criteria. The average ICC of each group ranged from 0.390 to 0.863, and the overall average was 0.659. There was no significant difference in inter-rater reliability according to the number of members in the group or the timing of the evaluation within the course.

Conclusion: The peer evaluation instrument that was used in the TBL classes was valid and reliable. Providing evaluation criteria and rules seemed to improve the validity and reliability of the instrument.

Key Words: Peer evaluation, Reliability, Validity, Team-based learning

Introduction

Team-based learning (TBL) is a well-defined instructional strategy that is increasingly employed in medical education because of its potential to promote active learning without requiring many faculty members [1]. TBL provides frequent opportunities for peers to enhance learning as teammates talk and listen to one another to arrive at consensus decisions. It also fosters individual and group accountability as small groups of students work together to solve problems and to answer questions [2]. To ensure accountability for group work, learners are asked to assess the contributions of peers within their group [3]. Peer evaluation methods in TBL have been constructed in many ways; however, limited

Received: October 31, 2017 • Revised: December 4, 2017 • Accepted: December 5, 2017

Corresponding Author: Wan Beom Park (https://orcid.org/0000-0003-0022-9625)
Department of Internal Medicine, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 03080, Korea
Tel: +82.2.740.8143 Fax: +82.2.745.6373 email: wbpark1@snu.ac.kr

Korean J Med Educ 2018 Mar; 30(1): 23-29.
https://doi.org/10.3946/kjme.2018.78
eISSN: 2005-7288

data exist regarding the validity and reliability of these instruments, especially in medical education [4].

Previous studies have examined the validity of peer evaluation instruments mostly by comparing peer evaluation scores with tutor evaluation scores or test grades [5-9]. However, tutors are not always available to supervise every small group activity, and tests usually measure knowledge instead of student contributions. Thus, the validity assessment of a peer evaluation method should examine whether the instrument actually measures a student's contributions in the right way. The other psychometric issue of such instruments is their reliability. If peer evaluation methods are to be considered reliable, one would expect that students who contribute the most to their groups should consistently get above-average peer evaluation scores from their peers, and low contributors should consistently get below-average peer evaluation scores (inter-rater reliability) [10]. The minimum number of members in each group to achieve the appropriate level of inter- rater reliability is also another issue regarding the reliability of peer evaluation instruments.

The purpose of this study is to assess the validity and reliability of a peer evaluation instrument that was used in TBL classes in a single medical school. This study examined whether the instrument actually measured students' contributions and whether the students who contributed the most to their groups consistently received higher peer evaluation scores from their peers and low contributors consistently got lower peer evaluation scores. This study also assessed whether the reliability of the instrument was affected by the number of students in a group or by the timing of the TBL class during the course.

Methods

At Seoul National University College of Medicine in South Korea, system-based integrated course starts from the fourth quarter of the first year. To promote active learning, TBL is frequently conducted throughout the course. A total of 146 students in the class of 2020 were divided serially into 18 groups by their student number for the TBL classes. However, five students dropped out before the beginning of the fourth quarter, and there were 141 students left. Eventually, five groups were composed of seven students, two groups were composed of nine students, and the other 11 groups were composed of eight students. The members of the groups were fixed until the end of the course. During the fourth quarter of the 2016 academic year, there were a total of 11 TBL classes. After every TBL class, the students were asked to evaluate their peers in the group. There were five criteria for the peer evaluation which were provided to the students. (1) Did the peer prepare enough for the class? (2) Did the peer actively participate in the group discussion? (3) Did the peer contribute to the group activity? (4) Did the peer respect others in the group? (5) Did the peer show sincerity during the class? The students were asked to rate their peers with an average of 10 points for each person; however, they had to rate at least one person above 11 points, and at least one person below 9 points. They were also asked to write a short comment for the person whom they gave the highest and lowest points in the group describing the reason for the high or low score. The students filled out the paper-based peer evaluation form by themselves after they left the classroom and placed them in a collection box the next morning. The evaluation of TBL classes was composed of iRAT (individual readiness assurance test) scores, tRAT (team readiness assurance test) scores, and peer evaluation scores. The proportion of peer evaluation scores was about 5% of the total course evaluation scores.

To assess the validity of the peer evaluation instrument, we analyzed the written comments for the highest and lowest achievers by the peer evaluation scores and examined whether the peer evaluation was conducted appropriately based on the criteria that the students were given. There were a total of 1,548 peer evaluation results, meaning that, there were 1,548 possible comments to the highest achievers and 1,548 possible comments to the lowest achievers. Two researchers in the study team independently reviewed all the comments and categorized them into groups according to the key concept of each comment. The comments which were categorized differently by each researcher were collected and reviewed again by both of them. The researchers discussed about each comment to reach a consensus.

The reliability of the instrument was assessed by examining the agreement among the performance ratings within each group of students via intraclass correlation coefficient (ICC) analysis. First, we examined whether the ICC differed among the groups. Second, we examined whether the ICC varied by the number of students in each group. Third, we examined whether the ICC changed throughout the course. Analysis of variance (ANOVA) was conducted to analyze differences among the ICCs. IBM SPSS ver. 23.0 (IBM Corp., Armonk, USA) was used for ICC and ANOVA analysis. The study was approved by the Institutional Review Board of Seoul National University College of Medicine and Seoul National University Hospital (IRB No. 1704-154-849).

Results

Among the total of 1,548 possible comments for each set, there were 1,317 comments for the highest achievers and 1,313 comments for the lowest achievers. Among the 1,317 comments for the highest achievers, 1,233 comments were positive and reasonable, while the other 85 comments were not specific and understandable. Among the positive comments, 901 comments were compatible with the criteria provided to the students, while the other 331 comments were not (Table 1). During the 11 TBL classes, the proportion of comments that were compatible with the criteria provided to the students increased moderately, while the proportion of comments that were not compatible decreased consistently (Table 2). Among the 1,313 comments for the lowest achievers,

Table 1. Composition of the Comments for the Highest Achievers

Comments	%
Compatible with the criteria	58.3
Active participation	41.9
Contribution to the team	9.3
Sincere attitude	3.3
Well preparedness	2.8
Respect for others	1.0
Not compatible with the criteria	21.4
Not reasonable	5.4
No comments	14.9
Total	100.0

Table 2. Characteristics of Podcasts Described

Variable	Class											Pearson's chi-square	p-value
	1	2	3	4	5	6	7	8	9	10	11		
Compatible with the criteria	51.0	53.3	58.2	51.8	58.6	56.4	60.7	68.8	65.2	58.8	57.8	74.36	<0.001
Not compatible with the criteria	34.3	30.0	24.1	31.2	21.4	23.6	18.6	11.3	12.8	11.8	14.8		
Not reasonable	2.1	4.7	5.0	5.0	7.9	7.1	7.1	5.7	5.7	5.9	4.4		
No comments	12.6	12.0	12.8	12.1	12.1	12.9	13.6	14.2	16.3	23.5	23.0		
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0		

1,207 comments were negative and reasonable, while the other 106 comments were not specific and understandable. All the negative comments were compatible with the criteria for evaluation (Table 3).

The average ICCs of each group ranged from 0.390 to 0.863, and the overall average was 0.659. There were significant differences among the average ICCs across groups ($p<0.001$). The average ICC of groups composed of seven students was 0.694, that of groups of eight students was 0.629, and that of groups of nine students was 0.783 (Table 4). There was no significant difference in the average ICC according to the number of students in the group ($p=0.075$).

The average ICC of each class ranged from 0.574 to 0.817, and there was no significant difference in the average ICC of each class during the course ($p=0.193$).

Table 3. Composition of the Comments for the Lowest Achievers

Comments	%
Compatible with the criteria	80.0
Passive participation	62.1
Less contribution to the team	4.0
Insincere attitude	9.2
Poor preparedness	1.8
Less respect for others	0.8
Not compatible with the criteria	0.0
Not reasonable	4.8
No comments	15.2
Total	100.0

When we divided the classes into three serial periods corresponding to their timing within the course (beginning, middle, and end), the average ICC of each period was 0.710, 0.632, and 0.629, respectively (Table 5). There was likewise no significant difference in the average ICC across the three periods ($p=0.090$).

Discussion

Most of the students provided reasonable and understandable comments for the highest and lowest achievers within their group, and most of those comments were compatible with the evaluation criteria that were given to the students. However, approximately one-fifth of the comments for the highest achievers were not compatible with the criteria. Some students mentioned that the high achievers exhibited good leadership, expressed creative and essential ideas, and gave well-organized presentations. There could be several reasons for this mismatch between the criteria and the comments. Students still might not have been familiar with the evaluation criteria at the beginning of the TBL classes. Indeed, the proportion of comments that were compatible with the criteria increased moderately, while the proportion of comments that were not compatible decreased con-

Table 4. The Average ICC of Groups by the Number of Members

No. of members in each group	No. of groups	Average ICC	SD	F-value	p-value
7	5	0.694	0.253	2.627	0.075
8	11	0.629	0.253		
9	2	0.738	0.161		

ICC: Intraclass correlation coefficient, SD: Standard deviation.

Table 5. The Average ICC of Each Period of Classes within the Course

Period	No. of classes	Average ICC	SD	F-value	p-value
Beginning	4	0.710	0.224	2.441	0.090
Middle	3	0.632	0.286		
End	4	0.629	0.247		

ICC: Intraclass correlation coefficient, SD: Standard deviation.

sistently during the course. However, all the comments for the lowest performers were compatible with the criteria throughout the course. Thus, it would be more reasonable to infer that the criteria provided to the students did not fully cover the aspects of the highest achievers in the TBL class. In that case, it would be better to modify the evaluation criteria to improve the validity of the instrument.

The total average ICC was 0.659, which is an acceptable level compared to previous studies on the inter-rater reliability of peer evaluation methods [11-13]. Several factors may have contributed to this result. First, in this study, the students were provided the evaluation criteria to guide them in the process of peer evaluation. Previous studies have shown that providing evaluation criteria to the students improved the reliability of the peer evaluation method [14]. Second, the students were also asked to rate at least one person in the group above 11 points, and at least one person below 9 points. This ensured discrimination among the peer evaluation scores of the students, which eventually increased the inter-rater reliability of the instrument. It is known that students, especially in medical school, feel uncomfortable giving their peers different scores, so that they tend to give similar scores to their peers if there are no other rules or regulations [15].

The average ICC of each group ranged from 0.390 to 0.863, and significant variance was present in the average ICC across groups. However, no significant difference in the average ICC was found according to the number of students in the group. A previous study showed that the ICC increased when the number of students in the group was increased from four or five to six in TBL classes [4]. From our study, we might infer that if there are more than six students in a group, increasing the number of students in the group does not significantly improve the reliability of the peer evaluation instrument. Thus, it seems that six to seven students might be the most appropriate number for a group in TBL classes to facilitate interaction among the students while not compromising the reliability of the instrument.

No significant difference was found in the average ICC of each class during the course. When we divided the classes into three serial periods corresponding to their timing in the course, the ICCs of the beginning classes were slightly higher than those of the classes in the middle and the end period, although this difference was not statistically significant. In this study we were not able to determine why the ICC slightly decreased from the beginning of the course to the middle. One possibility is that the students might have tried to give other students high scores in the middle and the end period of classes than the students whom they already gave the highest scores in the beginning period. This phenomenon is known as “gaming the system,” which refers to a tendency for students to try to even out the peer evaluation scores during the TBL classes [16]. Fortunately, in this study, the ICC did not decrease further after the middle period of classes.

This study also has some limitations. First, this study assessed the validity and reliability of a single specific instrument used in a single institution. Because every evaluation instrument has its own psychometric characteristics, the results of this study may not be directly applied to other peer evaluation methods in other circumstances. Second, the students were distributed into groups only by their serial student number, regardless of their age, gender, previous academic achievements, or other characteristics. There may been several factors influencing the group dynamics that were not fully considered in this study. Finally, we were not able to further investigate the reason for the differences in inter-rater reliability among the groups.

In conclusion, the peer evaluation instrument that was used in the TBL classes in a single medical school was valid and reliable. Most of the students assessed their peers' group activity and contributions based on the evaluation criteria. The students who contributed the most consistently got higher peer evaluation scores from other students in the group. No significant differences in inter-rater reliability were found among the groups according to the number of members in the group or the timing of evaluation during the course. Providing evaluation criteria and rules to the students seems to have improved the validity and reliability of the instrument. Further study is needed to explore the underlying group dynamics and to improve the validity and reliability of peer evaluation instruments.

ORCID:
Hyun Bae Yoon: https://orcid.org/0000-0003-4367-5350;
Wan Beom Park: https://orcid.org/0000-0003-0022-9625;
Sun Jung Myung: https://orcid.org/0000-0001-7332-0126;
Sang Hui Moon: https://orcid.org/0000-0003-1102-8714;
Jun-Bean Park: https://orcid.org/0000-0003-4053-8713

Acknowledgements: None.

Funding: None.

Conflicts of interest: No potential conflict of interest relevant to this article was reported.

Author contributions: HY carried out the study, conducted the analysis, and drafted the paper. WP led the design and implementation of the study and reviewed the final version before submission. SM, SM, KP, JP participated in the study and drafted the paper together. All authors read and approved the final manuscript.

References

1. Thompson BM, Schneider VF, Haidet P, et al. Team-based learning at ten medical schools: two years later. Med Educ. 2007;41(3):250-257.
2. Koles PG, Stolfi A, Borges NJ, Nelson S, Parmelee DX. The impact of team-based learning on medical students' academic performance. Acad Med. 2010;85(11):1739-1745.
3. Huh S. How to administer the peer evaluation in team-based learning. Korean J Med Educ. 2012;24(4):359-361.
4. Wahawisan J, Salazar M, Walters R, Alkhateeb FM, Attarabeen O. Reliability assessment of a peer evaluation instrument in a team-based learning course. Pharm Pract (Granada). 2016;14(1):676.
5. Ferguson KJ, Kreiter CD. Assessing the relationship between peer and facilitator evaluations in case-based learning. Med Educ. 2007;41(9):906-908.
6. Johnston L, Miles L. Assessing contributions to group assignments. Assess Eval High Educ. 2004;29(6):751-768.
7. Steensels C, Leemens L, Buelens H, et al. Peer assessment: a valuable tool to differentiate between student contributions to group work? Pharm Educ. 2006;6(2): 111-118.
8. 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.
9. Yoo S, Lee K, Lee SH, et al. Peer assessment of small-group presentations by medical students and its implications. Korean J Med Educ. 2014;26(1):31-40.
10. Strumpel CA. Peer assessment in the team-based learning classroom [dissertation]. Vancouver, Canada: University of British Columbia; 2011.

11. Arnold L, Willoughby L, Calkins V, Gammon L, Eberhart G. Use of peer evaluation in the assessment of medical students. J Med Educ. 1981;56(1):35-42.

12. Davis JD. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. Obstet Gynecol. 2002;99(4):647-651.

13. Dijcks R, Prince KJ, van der Vleuten CP, Scherpbier AJ. Validity of objective tests towards peer-rated competence by students. Med Teach. 2003;25(3):273-276.

14. Falchikov N, Goldfinch J. Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. Rev Educ Res. 2000;70(3):287-322.

15. Pocock TM, Sanders T, Bundy C. The impact of teamwork in peer assessment: a qualitative analysis of a group exercise at a UK medical school. Biosci Educ. 2010;15(1):1-12.

16. Levine RE. Peer evaluation in team-based learning. In: Michaelsen LK, Parmlee DX, McMahon KK, Levine RE, eds. Team-Based Learning for Health Professions Education: A Guide to Using Small Groups for Improving Learning. Sterling, USA: Stylus Publishing LLC; 2007:103-116.