Peer Evaluation in a Clinical Clerkship: Students' Attitudes, Experiences, and Correlations With Traditional Assessments

Ruth E. Levine, M.D., P. Adam Kelly, Ph.D., M.B.A.
Tayfun Karakoc, M.D., Paul Haidet, M.D., M.P.H.

Objective: *The authors performed this study to determine whether clerkship peer evaluations, initiated as part of our "team-based learning" curriculum in 2002, correlated with other student performance measures, and to determine what qualities students rate in their peer evaluations.*

Method: *The authors correlated peer evaluation scores with other student performance measures and performed a qualitative examination of student comments to assess reasons students gave for giving high and low scores.*

Results: *Peer evaluation scores correlated modestly with the National Board of Medical Examiners' (NBME) subject test, in-class quiz, and clinical scores. Qualitative comments demonstrated that students made assessments based on three thematic areas: personal attributes, team contributions, and cognitive abilities.*

Conclusions: *Peers' evaluation scores modestly predict which students will perform well on other measures. However, there may be other qualities that are also important factors in peer evaluation. For example, most students value qualities of preparation and participation. Though students sometimes dislike peer evaluations, their assessments may enhance traditional course assessments and complement a comprehensive evaluation strategy.*

Academic Psychiatry 2007; 31:19–24

Received July 13, 2005; revised March 7, 2006; accepted June 6, 2006. Drs. Levine and Karakoc are affiliated with the Department of Psychiatry, University of Texas Medical Branch, Galveston, Texas. Drs. Kelly and Haidet are affiliated with The HOuseton Center for Quality of Care and Utilization Studies and Baylor College of Medicine and the Houston VA Medical Center, Houston, Texas. Address correspondence to Dr. Levine, 301 University Boulevard, Route 0193, Galveston, TX 77555-0193; rlevine@utmb.edu (e-mail).

Peer evaluation can be a valuable tool to enhance students' performance. Using peer evaluation to assess interpersonal and professional behaviors is becoming increasingly important in medicine. Proponents highlight its potential to foster insight, reinforce other evaluations, and predict future performance. Studies of peer evaluation demonstrate positive correlations with faculty evaluations (1–3) and written exam performance (4–7).

However, the literature to date paints a complex picture. In some settings, students believed that they benefited from peer evaluation; in others they resisted the process (8–16). Students who accepted the method believed that the quality of their work improved based on the feedback given (8, 10). In other studies, students who disliked peer evaluation believed that it interfered with their relationships with fellow students (14, 16). We performed this study to explore student behaviors and perceptions in a novel peer evaluation process that was employed as part of a "team-based learning" instructional strategy in a psychiatry clerkship.

Method

Course Description and Peer Evaluation Strategy

At the University of Texas Medical Branch in Galveston, we introduced peer evaluation into our psychiatry clerkship at the same time that we started a new instructional method known as team-based learning. In team-based learning, also called "team learning," students are assigned to teams of four to six learners and presented with prerequisite readings, "readiness assurance" quizzes (completed first individually, then as a team), and complex application activities that are sequentially discussed among and across teams in a common classroom. Team-based learning leads to learning not only the facts but how those facts are applied in realistic settings. By practicing applications of facts

in the setting of complex and ambiguous problems, students gain a better understanding of the concepts that underlie individual bits of information and are better able to integrate new facts into their existing knowledge. Team-based learning has been successfully incorporated into both the pre-clinical and clinical curriculum of a variety of medical schools, with documented favorable learning outcomes, including effective stimulation of out-of-class study, favorable student attitudes about the value of teams, high levels of in-classroom engagement, and knowledge-based performance comparable to and sometimes superior to performance associated with conventional didactics (17–19). (A thorough description of team-based learning can be found at www.tblcollaborative.org.) Appendix 1 presents details of our application of team-based learning; a full description of our course has been published previously (18).

We adopted a peer evaluation strategy that was developed specifically for the team-based learning method (20). Peer evaluation is a core component of team-based learning and is seen as essential for holding students accountable and preventing the phenomenon of "social loafing." This strategy requires each student to discriminate among his or her peers so that at least one teammate gets a better than average score and at least one teammate gets a lower than average score. We chose this method because it was designed to reward contributions to the team, consistent with a central focus of team-based learning. At the beginning of each clerkship cycle, we informed the students that they would be engaging in team-based learning and that they would be performing a peer evaluation at the end of the 6 weeks. We asked them to come to consensus regarding whether they wanted the peer evaluation component to contribute toward their overall grade, and 75% of clerkship cohorts (six out of eight) chose not to have the peer evaluation count toward their overall grade. We employed the same structure for the peer evaluation process regardless of whether or not the peer evaluation score contributed towards the clerkship grades.

During the last week of each clerkship cycle, each student was given a paper and pencil "peer evaluation" form. Students were asked to assign scores with the following instructions:

> Assign scores for peer evaluation of helping behavior that reflect how you really feel about the extent to which the other members of your group contributed to your learning and/or group's performance. This will be your only opportunity to reward members of your group who actually worked hard on your behalf. If you give everyone pretty much the same score, you will be hurting those who did the most and helping those who did the least.

Each student was given a supply of 10 points per teammate and instructed to assign the points to the other members of the team. The instructions required at least one teammate to receive a score of 11 or higher and at least one teammate to receive a score of 9 or lower. For example, in a six-member team, each student was given 50 points to divide and assign. Students were then asked to describe their rationale in assigning the highest and lowest scores. We informed the students that these anonymous comments would be used to provide feedback to students who requested it.

Data Collection

The questions that guided our inquiry were: 1) How well did student peer evaluations correlate with other performance assessments? 2) What kinds of qualities did students use to judge their peers? 3) What were the perceived pros and cons of this peer evaluation strategy? In the published literature on peer evaluation, the conceptual relationship between student performance and peers' ratings is a positive one. A negative relationship does not make sense conceptually (i.e., poor performers being evaluated highly by peers). Therefore, in our analysis we employed one-tailed tests of statistical significance (a positive correlation versus no correlation).

We collected data from 152 students who rotated through the psychiatry clerkship during the 2002–2003 academic year. We computed and examined the Pearson product-moment correlations between peer evaluation scores and the following student performance scores: National Board of Medical Examiners' (NBME) psychiatry subject test scores, clinical grades, and individual and group readiness assurance test scores. We also performed a qualitative analysis of written rationale statements from 150 of our 152 students (two elected not to provide written statements) for their assignment of high and low scores to their peers. For this analysis, two of us (RL, PH) independently read all of the comments and underwent three iterations of discussion and rereading to devise coding categories based on themes that emerged from these discussions. One of us then coded all of the comments, and the other then verified the code assignments. The two authors reached 100% agreement on the assignment of codes.

Results

Quantitative

Descriptive statistics for the student performance scores and peer evaluation mean scores are shown in Table 1 (a

peer evaluation mean score is the mean of the three to five individual scores each student received from his or her teammates). We examined the multivariate distributions of test scores and found no outliers. That is, while some students performed extremely high or low on certain tests, there were no students who performed at the extremes on all tests overall. Correlations between peer evaluation mean scores and other teacher assessments were positive, albeit modest in magnitude, as shown in Table 2. We expected that peer evaluation mean scores, reflective of students' assessments by peers *as individuals*, would correlate more strongly with other individual performance scores than with a group performance score (i.e., the group readiness assurance test [GRAT]). This expectation was borne out. Also, the correlations between peer evaluation mean scores and student performance scores, with the exception of the GRAT, were similar in magnitude to the correlations of the student performance scores with each other.

Qualitative

The 152 students in the study turned in 150 forms and wrote a total of 343 statements. One hundred eighty-three statements gave a reason(s) for assigning a highest score, and 160 statements gave a reason(s) for assigning a lowest score. Many statements included more than one reason. We assigned a total of 22 mutually exclusive codes to these individual reasons and grouped these codes into four thematic areas (Table 3). An overwhelming majority of statements dealt with students' preparation (under the theme "personal attributes") or the quality of contributions during team discussions (under the theme "team process").

Examples of statements about preparation include:

- "Was consistently very well prepared and would do a good job of explaining why he thought an answer was correct or incorrect" (rationale for high score)
- "Rarely prepared to discuss the reading assignments" (rationale for low score)

Examples of statements about quality of contributions include:

- "The student worked well with the group members and positively contributed to the group's discussion week in and week out" (rationale for high score)
- "Minimal input—went along with the rest of the group" (rationale for low score)

Statements regarding preparation and quality of contributions to team communication were mentioned 120 and 115 times, respectively, and significantly overshadowed other reasons for giving someone high or low points on their peer evaluation scores. In the thematic area of "personal attributes," "attitude" was the second most common reason mentioned and was mentioned only 16 times. Students also valued the importance of cognitive abilities and mentioned as a reason for giving a high score "being knowledgeable" (N = 30), "depth of thought and reasoning" (N = 20), and "contributions during group quizzes" (N = 15).

When students were asked reasons for giving a low score, they also frequently mentioned the importance of preparation and contributions to the team process. Yet, students were frequently hesitant to give negative feedback about another student. Rather, they chose to comment about the peer evaluation process. Overall, students made 72 unsolicited statements either critical of the peer evaluation process or to the effect that "everyone contributed equally so evaluation was unnecessary." For example, one student wrote, "I actually thought everyone contributed equally. Some were more talkative, but as far as answering questions, we all 'saved the group' an equal number of times. You should not have to score people on a curve."

In all, 56 comments were coded as "negative comments about the peer evaluation process," while 16 students gave comments that could be summarized by the statement, "everyone on our team did well."

Discussion

In this study, we examined a peer evaluation process from two perspectives. Our quantitative analysis explored

TABLE 1. Descriptive Statistics for Quantitative Data

	Mean (SD)	Minimum	Maximum
NBME Score	72.7 (8.0)	53.00	99.0
Clinic Score	91.1 (2.9)	78.00	99.0
IRAT Score	83.8 (8.6)	57.00	100.0
GRAT Score	96.2 (3.3)	83.00	100.0
Peer Evaluation Mean Score	10.1 (0.9)	7.25	13.0

NBME = National Board of Medical Examiners
IRAT = Individual Readiness Assurance Test
GRAT = Group Readiness Assurance Test

TABLE 2. Correlations of Peer Evaluation Scores With Other Student Performance Scores, N = 152

	NBME Score	Clinic Score	IRAT Score	GRAT Score
Clinic Score	0.2780			
p value*	0.0004			
IRAT Score	0.4900	0.3170		
p value	0.0001	0.0002		
GRAT Score	0.2340	0.1650	0.3510	
p value	0.0020	0.0210	0.0001	
Peer Evaluation Mean Score**	0.3720	0.2800	0.4130	0.034
p value	0.0001	0.0004	0.0001	0.337

* t tailed test
** Mean of three to five scores each student received from his/her peers
NBME = National Board of Medical Examiners
IRAT = Individual Readiness Assurance Test
GRAT = Group Readiness Assurance Test

students' peer evaluations from a teacher's perspective, in that it examined the correlations with traditional forms of student assessment (e.g., board scores). Our qualitative analysis approached peer evaluation from a more student-oriented perspective, in that we tried to understand the parameters that students employed as they assessed their peers. By situating this study in a team-based learning context and employing multiple perspectives, we gained new insights into the value of peer evaluation, especially as it applies to situations in which students are expected to learn either in parallel or in collaboration with other students. In medical education, this study is relevant for a variety of contexts, such as problem-based learning groups, clinical clerkships, and laboratories.

TABLE 3. Thematic Areas and Codes for Assigning Peer Evaluation Scores

Thematic Areas and Codes	Number of Statements
Team Process	
Acts on behalf of the team	8
Quality of contribution to team communication	115
Leadership	6
Acting as equals	3
Listening abilities	7
Mediation abilities	1
Personal Attributes	
Consistency	4
Initiative	5
Attitudes	16
Work ethic	3
On time	1
Thorough	1
Student's feelings about other student(s)	2
Role modeling	2
Reliability	2
Advance preparation	120
Cognitive Abilities	
Contributions during group quizzes	15
Creative thinking	1
Knowledgeable	30
Depth of thought and reasoning	20
Negative Comments on the Peer Evaluation Process	
General negative comments	56
"We were all good"	16

In our quantitative analysis, we observed positive correlations with measures of individual knowledge acquisition, such as clinical grades and NBME scores. However, these correlations were only modest. One possible explanation for this is that students' assessments were based on other attributes in addition to factual knowledge. Our qualitative inquiry provided support for this explanation in that it illuminated a number of other personal and team process attributes that students value when evaluating colleagues. In an era of increasing attention to such attributes in the form of professional competencies (21), the results of our multi-method inquiry suggest that the assessments of peers may be a valuable complement to traditional test-based assessment data. In addition, since students in a team-based learning context spend significant amounts of time working and learning together, they may provide a unique viewpoint that other interpersonal assessments (e.g., resident or attending evaluations) may not capture. Further study about the processes and outcomes of student-focused peer evaluation strategies is warranted.

We were intrigued by the many students who identified contributions to team communication as an important factor driving their peer assessments. We suspect that the team-based learning process itself influenced this result.

Team-based learning uses a grading system where team performance (e.g., group readiness test scores) significantly affects individual student grades. We were encouraged to see that students performing peer evaluation in this context readily identified contributions to the team process as an important parameter of assessment. Overall, we believe our grading system in which team performance was a significant component, and our peer evaluation system in which team performance was rewarded, reinforced our clerkship goals of promoting interpersonal and professional communication. This finding points to the importance of the interaction between the learning context (i.e., the grading system) and the peer evaluation activity toward reinforcing core goals of a course or curriculum. We suggest that future work on peer evaluation focus on such interactions between the peer evaluation process and the learning context in which it takes place.

In addition to our main study findings, another issue came to light: many students spontaneously shared their dislike for the peer evaluation process. Given such student concerns, is peer evaluation really worth the effort? We suggest that it is. Although the desire to encourage professionalism in medical trainees is a popular initiative in modern medical circles, concrete exercises to act on this desire are scarce. In our clerkship, we decided to address student discontent with our peer evaluation system by eliminating the requirement to grade in a discriminatory fashion. Peer evaluation is still required, and narrative comments are still encouraged. Since implementing this change, students have already begun expressing a higher level of satisfaction with the peer evaluation process. Whether this new method will be sufficiently rigorous to both reinforce student accountability and provide meaningful feedback will require further study.

This study has several limitations, starting first with our peer evaluation form. By asking for nondescript "reasons for giving high scores and low scores," we solicited feedback that was vague and nonspecific. A more structured form with prompts and categories for feedback might have yielded different results. Nevertheless, we employed the form developed specifically for team-based learning, not for research purposes. A further limitation was that neither of the authors who interpreted the qualitative data was blind to the values of the students, since both interacted with these students through either direct teaching or focus groups.

Conclusions

The peer evaluation scores in our course correlated positively with other, largely "knowledge-based" student performance scores, suggesting that students value teammates for the same qualities that help them to perform well on the wards and on cognitive tests. However, the correlations were modest in magnitude. A possible explanation for this, supported by the findings of our qualitative analysis of students' rationales, is that there are other, "noncognitive" qualities that are also important factors in peer evaluation. These noncognitive qualities may be the most important components of peer evaluation, since they are most likely to be correlated with interpersonal and professional skills. As we improve our methods of helping students to assess these skills in each other, we believe we will come closer to our goal of promoting professionalism.

APPENDIX 1. Description of the Team-Based Learning Component of Our Psychiatry Clerkship

Team-Based Learning Element	Specific Components
Description of Scope	8 hours over 6 weeks with 20 to 24 students
Team Formation	Teams of four to six students assigned by course director and remaining together throughout rotation. All sessions are completed in one classroom with all teams present
Readiness Assurance	Individual (IRAT) and group readiness assurance tests (GRAT) were conducted (closed-book) for all readings
Immediate Feedback	Faculty review GRAT answers after each test
Problem-Solving Sequencing	Challenging multiple-choice problems are given to the teams, with intra-team discussion followed by inter-team discussion. Problem-solving activities alternate with readiness assurance tests
Format of Problem-Solving Activities	All teams work on the same problem; problems are formatted as clinically oriented multiple-choice questions. Teams simultaneously report their solutions and the teacher fosters a discussion of differences across groups
Incentive Structure	Students graded on IRAT, GRAT and peer evaluation. Team-based learning grade is 15% of overall clerkship grade

IRAT = Individual Readiness Assurance Test
GRAT = Group Readiness Assurance Test

References

1. Arnold L, Willoughby L, Calkins V, et al: Use of peer evaluation in the assessment of medical students. J Med Educ 1981; 56:35–42
2. Von Rosendaal GMA, Jennett PA: Comparing peer and faculty evaluations in an internal medicine residency. Acad Med 1994; 69:299–303
3. Sullivan ME, Hitchcock MA, Dunnington GL: Peer and self-assessment during problem-based tutorials. Am J Surg 1999; 177:266–269
4. Ramsey PG, Wenrich MD, Carline JD, et al: Use of peer ratings to evaluate physician performance. JAMA 1993; 269: 1655–1660
5. Eva KW: Assessing tutorial-based assessment. Adv H Sci Educ 2001; 6:243–257
6. Norcini JJ: Peer assessment of competence. Med Educ 2003; 37:539–543
7. Cheng W, Warren M: Making a difference: Using peers to assess individual students' contributions to a group project. Teach Higher Educ 2000; 5:243–255
8. Heylings DJ, Stefani LAJ: Peer assessment feedback marking in a large medical anatomy class. Med Educ 1997; 31: 281–286
9. Thomas PA, Gebo KA, Hellmann DB: A pilot study of peer review in residency training. J Gen Intern Med 1999; 14:551–554
10. Magzoub ME, Schmidt HG, Dolmans D: Assessing students in community settings: the role of peer evaluation. Adv Health Sci Educ Theory Pract 1998; 3:3–13
11. Vuorinen R, Tarkka MT, Meretoja R: Peer evaluation in nurses' professional development: a pilot study to investigate the issues. J Clin Nurs 2000; 9:273–281
12. Ramsey PG, Carline JD, Blank LL, et al: Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. Acad Med 1996; 71:364–370
13. Wendling A, Hoekstra L: Interactive peer review: an innovative resident evaluation tool. Fam Med 2002; 34:738–743
14. Van Rosendaal GM, Jennett PA: Resistance to peer evaluations in an internal medicine residency. Acad Med 1992; 67:63
15. Reiter HI, Eva KW, Hatala RM, et al: Self and peer assessment in tutorials: application of a relative-ranking model. Acad Med 2002; 77:1134–1139
16. Greenbaum DS, Hoban JD: Teaching peer review at Michigan State University. J Med Educ 1976; 51:392–395
17. Searle NS, Haidet P, Kelly PA, et al: Team learning in medical education: initial experiences at 10 institutions. Acad Med 2003; 78(suppl 10):555–558
18. Levine RE, O'Boyle M, Haidet P, et al: Transforming a clinical clerkship through team learning. Teach Learn Med 2004; 16: 270–275
19. Michaelsen LK, Knight AB, Fink LD (eds): Team-Based Learning: A Transformative Use of Small Groups in College Teaching. Sterling, Va, Stylus Publishing, 2004
20. Michaelsen LK, Fink LD (eds): Calculating peer evaluation scores, in Team-Based Learning: A Transformative Use of Small Groups in College Teaching. Sterling, Va, Stylus Publishing, 2004, pp 241–248
21. Association of American Medical Colleges: Learning objectives for medical student education: guidelines for medical schools: report I of the Medical School Objectives Project. Acad Med 1999; 74:13–18