EVALUATING MEDICAL STUDENT CLINICAL SKILL PERFORMANCE
Relationships Among Self, Peer, and Expert Ratings

JUDITH G. CALHOUN
JAMES O. WOOLLISCROFT
JOCELYN D. TEN HAKEN
FREDRIC M. WOLF
WAYNE K. DAVIS

Departments of Postgraduate Medicine, Health Professions Education, and Internal Medicine
University of Michigan Medical School

Second-year medical students (N = 187) evaluated their own videotaped performances of one of eight randomly assigned physical assessment examinations. The videotaped performance was one component of an introduction to clinical sciences course evaluation. Performance ratings were also obtained from two peers—one who served as the patient and the other who served as the camera person for the evaluation—and one expert. All rater groups used the same behaviorally anchored evaluation checklist of the key techniques and sequences identified for each examination. High Pearson product-moment correlations were obtained between (1) the two peer ratings for four of the examinations and (2) self and peer ratings for the other four examinations. Repeated measures analysis of variance revealed significant differences among the four types of raters for all but one of the eight different examinations. Implications for future evaluation methodologies and curricular implementation of peer assessment are discussed.

Evaluation & the Health Professions, Vol. 11 No. 2, June 1988 201-212

Peer assessment has been shown to be an effective strategy in a variety of assessment contexts, including private industry, the military, and education. Nevertheless, there has been reluctance to adopt it as a basis for evaluative decision making due to commonly held views that such assessments (1) are little more than popularity contests and (2) do not provide any information which is different from, or psychometrically superior to, that which exists from other sources (Kane and Lawler, 1978).

Despite the recognition that peer assessment provides career-long responsibility for oneself and colleagues (Engebretsen, 1977; Linn et al., 1975) and is a critical ingredient in the process of continuing medical education (Linn et al., 1975; Peterson, 1972), the technique is not widely used in medical schools (Engebretsen, 1977). Instead, as Engebretsen points out, evaluation of the physician-in-training has traditionally rested with the educator, who is commonly pictured as authoritarian and paternalistic, though perhaps benevolent. As a result, he states, it is not surprising that the medical education process has produced a generation of physicians who recoil at the thought of peer review, with which they have little or no experience. He further proposes that if the long-range goal is to affect the current health care system by influencing the behavior of future physicians, peer evaluation will have to be integrated into the educational process.

Most of the studies published on the use of peer assessment in medical education have dealt with the technique from a research perspective. Only three studies could be identified that focused upon peer evaluations from an administrative setting as part of the instructional process or formal assessment of medical students. Engebretsen (1977) reported that when peer review was included in a new system of governance for the residency program, faculty interest, responsibility, and involvement in the education-evaluation process were increased. In their study of the measurement properties of peer evaluations conducted during the undergraduate curriculum years, Arnold et al. (1981) found that peer ratings given as part of a promotion process were internally consistent, unbiased, and valid. They also reported that medical students tend to rate their peers in a global fashion rather than

along the discrete dimensions contained in an evaluation instrument. In contrast to prior studies, however, they did not find as high a level of interrater agreement among the peer ratings or that peers provided unique information about student performance. Pepe et al. (1980) were the only researchers identified who addressed the students' reactions to using peer assessment. They anecdotally reported that students valued the peer review and feedback they received while practicing their interviewing, data-recording, and clinical problem-solving skills.

The purpose of this study was to investigate the value of incorporating peer ratings in the teaching and evaluation of physical assessment skills. Specific areas of investigation included (1) students' ability to perceive and interpret accurately the salient aspects of their own and their peers' clinical skills performance, (2) the extent to which self, peer, and expert skill evaluations were similar, (3) the extent to which peer roles affected the peer evaluations of the same students' performance, (4) students' reactions to assessing their own and their classmates' performances as part of a formal evaluation process, and (5) the use of peer ratings as a routine part of student clinical skill evaluation.

METHODS

Medical students (N = 236) enrolled in the second-year introduction to clinical sciences course served as the study group for the project. Instructional strategies for the physical examination skill sequence of this course included the following: (1) lectures emphasizing physical assessment techniques and sequencing, rather than analysis of findings and diagnosis, (2) videotaped demonstrations prepared by faculty of each physical assessment examination, (3) weekly clinical laboratory sessions for student practice of the different physical assessment examinations under faculty guidance, (4) textbook assignments, and (5) directive outlines detailing the sequence and techniques necessary for performing each aspect of the physical assessment examination.

Each of the 236 students was randomly assigned to a small

group of three for the purpose of evaluating the student's physical examination skills at the end of the course. The students were informed at the beginning of the course that they would be required to submit a videotape of their performance of one randomly assigned physical examination for evaluation at the end of the skills sequence. The students were also encouraged to attend the weekly clinical skills practice sessions and to use the videotape resources available in the school's learning resource center while practicing their skills.

There were 15 physical assessment examinations covered in this skills sequence. However, only 8 of the 15 physical examinations were deemed by the faculty and project staff to be acceptable for evaluation by the videotape format. These 8 were the general physical, abdominal, cardiac, musculoskeletal, neurologic, ophthalmologic, otorhinolaryngologic, and pulmonary examinations. The students, however, were not informed that any of the examinations would be deleted from the evaluation.

Each student in the evaluation group performed three different roles: examiner, patient, and camera person. The student groups were randomly assigned to one of nine evaluation periods over a two-week period near the end of the fall semester. At each scheduled evaluation period, the assigned students met with an instructor to (1) review the final instructions for the evaluation, (2) discuss questions and concerns, and (3) receive the specific physical assessment examination to be performed. Each member of the group was randomly assigned a different examination to perform. Based upon preliminary trials, each group was given a guideline of approximately four hours for (1) practicing the assigned physical examination until they were confident of their performance, (2) videotaping their performance, with the option of retaping sequences with which they were not satisfied, and (3) evaluating their own performance as well as those of the other two group members. The students were reminded that all regulations for taking examinations at the school applied for the evaluation, and that once in the clinical skills laboratory for the test they were not to refer to notes or discuss the examination sequence or techniques with one another.

Behaviorally anchored evaluation checklists for each of the physical assessment examinations were developed based upon (1) a task analysis of the skills demonstrated in the faculty videotapes, (2) a review of the literature and checklists developed by other institutions, and (3) faculty input regarding the key techniques and sequences deemed of importance for inclusion (see Table 1). Although the students were not given the checklist prior to the evaluation, they were given expanded descriptive outlines that included every criterion item included on the checklist. The expert evaluators for the project were chosen from the faculty or project staff who were closely associated with the task analysis and development of the course materials and resources for each examination. Four faculty and four project staff, including education specialists and nurses, served as the experts.

At the end of each evaluation session, the students reviewed the videotapes for their group and used the checklist to independently rate their own performance and that of their peers. The students were also asked to provide input regarding their opinions of the positive and negative aspects of the experience. Each faculty/staff evaluator was assigned only one of the physical examinations; thus all students performing the cardiac examination were assessed by the same evaluator, all pulmonary examinations were assessed by another evaluator, and so on. The designated evaluator for each physical assessment examination also used the checklist to evaluate the respective videotaped performances. The videotapes for those students who failed the evaluation were reviewed by the course director.

To determine if peer evaluation differed based on the role played during the evaluation, the peer evaluations were divided into peer-patient and peer-camera person categories. Pearson product-moment correlation coefficients were computed to identify the relationships among the self, peer, and expert ratings. Repeated measures analysis of variance was used to test for differences among the mean self, peer-patient, peer-camera person, and expert ratings of performance. Performances were examined separately for each physical examination because of the differences in exam length, exam content, and faculty

TABLE 1
Abdominal Examination Evaluation Checklist
(Example of Partial Checklist)

	Performance		
	Correct	Incorrect	Not Done
Inspection			
1. Patient suppine on exam table.	_____	_____	_____
2. Undrapes abdomen from above xiphoid process to symphysis pubis.	_____	_____	_____
3. Examiner begins assessment standing at patient's right.	_____	_____	_____
4. Visually inspects entire abdomen.	_____	_____	_____
Auscultation			
5. Uses (diaphragm) of stethoscope.	_____	_____	_____
6. Auscultates RUQ	_____	_____	_____
7. Auscultates RLQ	_____	_____	_____
8. Auscultates LLQ	_____	_____	_____
9. Auscultates LUQ	_____	_____	_____
10. Auscultates epigastric region.	_____	_____	_____
11. Auscultates abdominal aorta.	_____	_____	_____
Percussion			
12. Places one hand on the area to be examined and with one or two fingers of the other hand, strikes the hand resting on abdomen.	_____	_____	_____
13. Percusses RUQ	_____	_____	_____
14. Percusses liver height along the mid clavicular line.	_____	_____	_____
15. Percusses RLQ	_____	_____	_____
16. Percusses LLQ	_____	_____	_____
17. Percusses LUQ	_____	_____	_____
18. Percusses epigastric region	_____	_____	_____

evaluations. In addition, Scheffé a posteriori comparisons were conducted to identify specific pairwise differences among the four types of raters.

RESULTS

Complete data were available for 187 (79%) of the students. Because of the logistics of conducting such a large and intricate evaluation project, some of the peer roles (i.e., camera or patient) were not recorded and could not be identified for statistical analysis.

Table 2 presents the correlation coefficients for the self, peer-patient, peer-camera person, and expert ratings for each of the eight physical assessment examinations. The correlations for all examinations ranged from .24 to .90. For four of the examinations (cardiac, neurologic, ophthalmologic, and pulmonary), the strongest correlations were obtained between the two peer ratings. In contrast, the strongest relationships for the other four examinations were between the self and one or both of the peers.

The least agreement was found between the peer-patient and the expert ratings on six of the eight exams (musculoskeletal, neurological, abdominal, cardiac, general physical, and ENT). For the ophthalmologic examination, the lowest correlation was found between the self and the expert, whereas for the pulmonary examination, the lowest correlation was found between the camera person and the expert.

The mean percentage performance ratings for each type of rater by type of examination are listed in Table 3. Significant rating differences among the four types of raters were found for all but the abdominal and cardiac examinations. The self and peer ratings were quite similar for all of the examinations. It is interesting to note that the students rated themselves the same or slightly higher than their peers did on four of the examinations (abdominal, ENT, general physical, and musculoskeletal). Only on the cardiac assessment were the self ratings lower than both of the peer ratings. However, none of these self-peer differences was statistically significant. In contrast, the results of the Scheffé analysis on the exams with significant repeated measures revealed that the experts rated the students significantly ($p < .01$) lower than the students rated themselves or their peers rated them.

Student reactions to the experience immediately after com-

TABLE 2
Relationships Among Peer and Expert Performance Rating by Type of Physical Assessment Examination

		Mean r			Mean r		Mean r
		Self with:			Peer Patient with:		Peer Camera with:
	N	Peer Patient	Peer Camera	Expert	Peer Camera	Expert	Expert
Abdominal	28	.51**	.68**	.56**	.60**	.42*	.52**
Cardiac	27	.80**	.87**	.80**	.88**	.73**	.82**
ENT	23	.90**	.90**	.87**	.89**	.76**	.87**
General Physical	22	.76**	.65**	.71**	.71**	.57**	.62**
Musculoskeletal	24	.53**	.62**	.48*	.51*	.24	.40
Neurologic	20	.69**	.78**	.44	.90**	.35	.48*
Ophthalmologic	20	.50*	.61**	.37	.65**	.50*	.46*
Pulmonary	23	.59**	.68**	.59**	.78**	.58**	.55**
Combined	187	.73**	.78**	.64**	.80**	.59**	.64**

*$p < .05$; **$p < .01$.

TABLE 3

Mean Percentage Performance Ratings by Rater and Type of Physical Assessment Examination

	N	# of Items	Self		Peer Patient		Peer Camera		Expert		F
			Mean %	S.D.	Mean %	S.D.	Mean %	S.D.	Mean %	S.D.	
Abdominal	28	15	93.6	7.1	93.1	8.2	92.1	8.9	90.0	12.2	1.64
Cardiac	27	24	80.9	18.5	82.6	18.7	81.9	16.4	78.7	19.0	1.20
ENT	23	24	81.2	13.3	80.1	16.7	78.3	15.7	71.4	17.2	11.94*
General Physical	22	69	80.4	10.3	78.7	10.6	80.0	11.0	73.9	11.5	4.95*
Musculoskeletal	24	73	84.4	8.2	84.2	8.0	83.5	7.8	62.8	11.3	59.91*
Neurologic	20	35	83.7	8.5	83.9	8.8	80.6	10.6	77.0	13.6	4.16*
Ophthalmologic	20	17	82.6	13.7	86.5	9.9	82.6	12.3	69.4	9.5	17.03*
Pulmonary	23	11	87.7	15.4	85.0	13.9	89.3	14.7	75.5	17.2	9.80*
Combined	187		84.5	12.5	84.4	12.6	83.8	12.6	75.3	14.4	

*$p < .05$.

pleting the videotaping and completing their ratings were quite positive. They reported that overall the experience was valuable, enjoyable, well organized, and stimulated them to learn more and practice and review their skills. In addition, they felt that self and peer evaluations were extremely helpful and informative. Furthermore, the functioning and operation of the equipment did not interfere with the conduct of the evaluation. Negative comments addressed the stress associated with the evaluation, the timing of the evaluation at the same time as other final exams, the noise due to so many people in the clinical skills laboratory at one time, and the fact that some received much longer assessments to perform than others (i.e., general physical).

DISCUSSION

In contrast to prior research findings (Linn et al., 1975; Kubany, 1957; Hammond and Kern, 1959), the results of this study indicate that second-year medical students do not accurately assess their own or their peers' skills performance. Although the students' self ratings were quite similar to those of their peers, a result that supports the findings of Hammond and Kern (1959), these ratings differed significantly from expert assessment of the performance. Methodologically it could be argued that perhaps the experts were in error. However, given the prior clinical skills and teaching and evaluation experience of the experts, as well as their familiarity with the course materials and student performance of each criterion item included in the evaluation, the expert designation is adequately justified. Furthermore, the results regarding the statistically significant differences among the expert and student ratings, both self and peer, were consistent for seven of the eight physical assessment examinations. Even for the one exception to this finding, the abdominal examination, the expert rating was lower than the self and peer ratings.

A detailed review of the three prior studies (Linn et al., 1975; Kubany, 1957; Hammond and Kern, 1959), which found that medical students could reliably assess or rank their classmates,

revealed that third- and fourth-year clerkship students served as the subjects. Furthermore, none of these studies specifically addressed physical assessment examination performance. Only the study by Linn et al. (1975) included an attribute dealing with some measure of clinical skill behavior, labeled "technical ability" by these authors. It is proposed that the differences in results could be attributed to the level of the student involved in the peer assessment and the nature of the performance area. Carter (1962) points out that psychologists will probably agree that factors other than one's own ability to perform a task are involved in the capacity to rate the performance of others. It appears that this point applies to medical students assessing the clinical skill of their peers. Second-year medical students who are newly acquiring physical assessment knowledge and skills do not seem to have progressed to the evaluative level of cognitive functioning in this area.

The fact that the students in this study did not rate themselves lower than their peers rated their performance also differs from prior findings (Linn et al., 1975). Nevertheless, the result matches that of Hammond and Kern (1959), who found that fourth-year medical students agreed among themselves to a fairly high degree in their peer judgments. As noted by both the Arnold et al. (1981) and Hammond and Kern (1959) studies, it appears that medical students tend to base their evaluations on a global overview or clusters of attributes of their peers rather than making clear-cut decisions on specific areas of performance. A review of the student ratings for the discrete dimensions used in the evaluation checklists for this study indicates support for this aspect of student assessment.

Also of note were the a priori assumptions that the authors held regarding the expected differences due to the various examinations, as well as the two peer roles. These assumptions proved unfounded when the results indicated no significant differences across the eight exams or between the peer-camera and peer-patient evaluations. Therefore, future research endeavors in self and peer assessment by these authors will be based on these similarities across exams and peer raters.

This study represents one of the few in medical education that specifically analyzes peer assessment in an instructional and evaluative context. Other studies will have to be conducted to determine the extent to which the findings can be generalized. However, the study has provided baseline information for future evaluation methodologies, curricular implementation of peer assessment, and research regarding the evaluative maturation of medical students as they progress through the clinical years. As for other skills, students have to be introduced to the concept of peer review and evaluation as well as allowed to experience and practice the activity before they will become comfortable enough to incorporate it into their professional behaviors.

REFERENCES

ARNOLD, L., L. WILLOUGHBY, V. CALKINS, L. GAMMON, and G. EBERHARDT (1981) "Use of peer evaluation in the assessment of medical students." J. of Medical Education 36 (January): 35-42.

CARTER, H. D. (1962) "How reliable are good oral examinations?" California J. of Educ. Research 13 (September): 147-153.

ENGEBRETSEN, B. (1977) "Peer review in graduate education." New England J. of Medicine 296 (May): 1230-1231.

HAMMOND, K. R. and F. KERN [eds.] (1959) Teaching Comprehensive Medical Care. Cambridge, MA: Harvard Univ. Press.

KANE, J. S. and E. E. LAWLER (1978) "Methods of peer assessment." Psych. Bull. 85 (March): 555-586.

KUBANY, A. J. (1957) "Use of sociometric peer nominations in medical education." J. of Applied Psychology 41 (June): 389-394.

LINN, B. S., M. AROSTEGUI, and R. ZEPPA (1975) "Performance rating scale for peer and self assessment." British J. of Medical Education 9 (June): 98-101.

PEPE, E. A., C. G. HODEL, and D. A. BOSSHART (1980) "Use of peers to teach interviewing and clinical problem-solving." J. of Medical Education 55 (September): 800.

PETERSON, P. (1972) "Teaching peer review." J. of the Amer. Medical Assn. 224 (May): 884-885.