

Assessing Medical Student Professionalism: An Analysis of a Peer Assessment

Scott Cottrell, EdD*, Sebastian Diaz, PhD, JD†,
Anne Cather, M.D‡, and James Shumway, PhD§

*Assistant Professor, Department of Community Medicine and the Office of Medical Education. West Virginia University

†Assistant Professor, School of Human Resources and Education. West Virginia University

‡Associate Dean, Student Services. West Virginia University

§Associate Dean, Medical Education. West Virginia University

Abstract:

Purpose: Students' professional development is an essential aim of medical school. The purpose of this paper is to report how an assessment of first-year medical students' professional behavior was designed and to investigate its measurement characteristics.

Methods: The assessment was implemented as a peer assessment of professional skills, which were delineated according to a formal professional code developed by our curriculum committee. During the last week of the Fall 2005 semester, the professionalism assessment was administered online to students in a problem-based learning course.

Results: The internal consistency of the assessment is adequate. The generalizability study found that raters nested within persons accounted for the majority of variance. While the inter-rater reliability is relatively low, using multiple raters may yield an acceptable estimate of the relative reliability.

Conclusions: The results suggest that this peer assessment is a practical assessment, evidenced by the 91% compliance rate. However, future research and modifications will be needed to address the variance of responses, helping to discriminate between "poor" and "good" observations of professionalism. In addition, multiple raters are required to supply reliable estimates of students' professional behavior. Coupling this evaluation with other professionalism evaluations may help reveal a more complete picture of students' professional behavior.

Keywords: Problem Based Learning, PBL, Evaluation, Ethics/professionalism, Peer Assessment

Medical students' professional development is an essential aim of medical school. Much has been written about the characteristics of the construct of professionalism. Herbert Swick¹, for example, has offered a comprehensive definition of professionalism: 1) subordinating one's self-interests to the interests of others, 2) adhering to high ethical and moral standards, 3) evincing core humanistic values, 4) incorporating self-reflection about one's actions, 5) exercising accountability, 6) dealing with high levels of complexity and uncertainty, 7) exhibiting a commitment to scholarship, 8) demonstrating a commitment to excellence, 9) responding to societal needs, and 10) reflecting a social contract with communities served. Competency-based curriculum literature has also prompted attention to promoting professional development. Frameworks, such as CanMEDS², the University of Dundee Three-Circle Model³, the ACGME Competencies⁴, and the

AAMC Medical Student Objectives Project⁵, have also described the knowledge, skills and attitudes that define professionalism. These competency-based frameworks have helped define the construct of professionalism and display it as a collection of measurable variables. As Louise Arnold⁶ explained in her review of the literature, approximately half of all medical schools have written explicit criteria and specific assessment methods to assess professional behavior.

Despite efforts to define professionalism, questions continue to surface about how professionalism should be incorporated into the medical school curriculum. Some schools are developing learning opportunities and assessments to target professionalism. For example, Tulane University created the Program for Professional Values and Ethics in Medical Education (PPVEME).⁷ The program brings together students, residents, and

faculty into learning teams to discuss issues ranging from integrity, communication, leadership, and service. Another example includes a gross anatomy course that is taught at the Mayo Clinic College of Medicine.⁸ As a part of the evaluation process, students are asked to complete professionalism evaluations on their peers and themselves. The course directors concluded that implementing professionalism evaluations early in the medical school curriculum was a valuable exercise.

The purpose of this paper is to report how an assessment was designed to gauge medical students' professional behavior in a problem-based learning course. The assessment was implemented as a peer assessment of professional skills, which were delineated according to a formal professional code developed by our curriculum committee. The problem based learning course requires students to demonstrate skills consistent with many of the characteristics of professionalism, such as demonstrating accountability and respectful behavior. Investigating the measurement characteristics of this peer evaluation will help determine whether it can distinguish students who may not be demonstrating appropriate professionalism skills. The results may also have implications on whether the peer evaluation can be applied in other areas of the medical school curriculum.

Method and Materials

Targeted Professionalism Characteristics -

An important consideration of the assessment was the definition and specificity of the construct (i.e., professionalism) we are attempting to measure.⁹ That is, we need to define the construct and determine whether our scale targets relatively broad or narrow professional skills. Given these important considerations, our medical school's Code of Professionalism was used to inform the development of the assessment. Using existing literature and frameworks (e.g., AAMC Medical Student Objectives Project), the code was developed by the medical school curriculum committee. We defined the construct of professionalism as student behaviors that are characterized across nine domains of professionalism, which are similar to those described by Swick. The nine domains include:

- Honesty and Integrity
- Accountability
- Responsibility
- Respectful and Nonjudgmental Behavior
- Compassion and Empathy
- Maturity
- Skillful Communication
- Confidentiality and Privacy in all patient affairs

- Self-directed learning and appraisal skills

Using these domains, nine questions were developed to ascertain a comprehensive picture of students' professional skills. To help evaluators gauge students' professional knowledge, attitudes and skills, each question uses scoring rubrics. These rubrics help the evaluator navigate through complex responses and tease out important characteristics of professional behavior across the nine domains.

Huba and Freed¹⁰ suggested primarily two steps that assisted the design of the scoring rubric. First, it is necessary to identify the criteria that set the stage for characterizing levels of achievement. For example, if the evaluator is asked to assess a student's ability to be accountable, then poor, average, and excellent performance must be distinguished. After exploring the qualities of each domain, a bipolar scale was created-with extreme ends of the rubric suggesting *too much* or *too little* demonstration of a particular skill. For example, a student may fail to complete tasks, miss classes, and exhibit little or no accountability to the group. Conversely, a student may be excessively dominate, controlling many of the group's decisions and processes. Given these extreme ends of the same domain (e.g., accountability), the preferred level of achievement was the middle of the continuum.

Second, the standards of performance at each achievement level by anchoring three levels of the rubric's continuum need to be identified. The value corresponding to a score of "0" was anchored as "Not Observed." The first, fourth, and seventh levels (1, 4, 7) were anchored with detailed descriptions, allowing the respondent to intuit levels 2, 3, 5, and 6, which were left unanchored (see Table 1). These levels, which are included in the columns of the rubric, describe the range of performance. For example, a student's demonstration of a particular domain can vary on 7 points and a "not observed column."

Implementation of the Assessment - During the first year of our medical school curriculum, multidisciplinary faculty facilitate a PBL learning experiences. All first-year medical students (111) complete the problem-based learning course. The PBL course consists of 14 groups; each includes one facilitator and approximately 8 students in a group. The course consists of 5 cases throughout the semester. Each case is divided into three sessions. During the first session of each case, students are given information that describes a patient's chief complaints, history, and physical symptoms. Students are asked to share and explore hypotheses of the patients' condition.

Table 1
Peer Assessment Items and Anchors for Scale Points 1, 4, and 7

Item	Scale Anchors		
	1	4	7
Honesty	Misrepresents one self or knowledge; falsifies data; doesn't admit mistakes	Honest in actions and words: doesn't lie, cheat, steal, or plagiarize: admits mistakes:	Honest to the point of insensitivity to others, tactless
Accountability	Doesn't complete tasks: misses appointments: avoids work	Punctual to class or duty; well prepared: willing to accept praise	Controlling: excessive fault finding: self-righteous
Responsibility	Doesn't comply with policies, rules and regulations: is not prompt, prepared or organized: makes	Reliable; trustworthy: takes ownership of assignments: seriously and diligently works on assigned tasks:	Inflexible: rule-bound to the point of obstruction or paralysis: afraid to act out of fear of committing an error:
Respectful	Doesn't realize limitations of own beliefs and perspectives: discourteous; belittling	Consistently civil and courteous to all: tolerates diversity: listens before acting	Excessive selflessness: overextends oneself; nonjudgmental to the point of inaction
Compassion	Little compassion for others; appears cold, heartless, indifferent	Respects & is aware of others' feeling: shows mindfulness & self-reflection	Loses objectivity by excessive desire to help: emotionally labile and unduly empathetic;
Maturity	Makes unsound decisions: doesn't manage time well: can't maintain personal or professional boundaries:	Shows personal growth; recognizes & correct mistakes; tries to improve self; manages relationships & conflicts	Puts others ahead of self to a fault: attempts to improve oneself to a fault; perfectionist
Communication	Unable to communicate at another's level of understanding; writes illegibly;	Effective use of oral, written & non-verbal skills: speaks with clarity to all; culturally appropriate skills;	Gives feedback when not solicited; constantly assumes role of conflict manager;
Confidentiality	Seeks patient information when unnecessary or inappropriate:	Maintains information appropriately: acts in accordance with known guidelines	Inappropriately assumes role as watchdog for patient confidentially and privacy violations
Self-directed learning	Accomplishes tasks with excessive assistance of others	Displays ability to be a life long learner; completes all evaluations	Dominant, overbearing, authoritative in team settings:

Students identify key learning issues, or questions about the material. The learning issues, which drive students' self-directed learning skills, are researched before the next session.

During the second session, students share and discuss the collected information that addresses the learning issues. Additional information about the patient is given, yielding more learning issues for the third, and final, PBL component. The last session begins with the presentation and discussion of the learning issues. Addressing learning issues helps students learn the basic science associated with the case and refine the hypotheses about the patient's problem, eventually leading to a course of action and a full discussion about the implications of the medical condition.

As students cooperatively share information, PBL aims to develop skills that are characteristic of professionalism. Students, for example, demonstrate interpersonal skills as case information is presented and critiqued. Each student must also share responsibility for the group, as they are dependent on collecting detailed and accurate information about the case.

During the last week of the Fall 2005 semester, the professionalism assessment was administered online to students. IRB approval was granted and student names remained anonymous. The results were not given to facilitators. In addition, the results did not affect the evaluation of students' performance in the PBL course.

Results

General Psychometric Properties of the Peer Professionalism Assessment - There were approximately 8 students in each of the 14 PBL groups. The students were asked to rate their peers, totaling to approximately 7 ratings for each student. Across the fourteen PBL groups, approximately 91% of the peer assessments were completed. Before performing any statistical analyses, the data were explored. An analysis of the responses from each rater revealed that one student obviously did not understand the bipolar scale. The student gave each subject a rating of 7, which suggests that each subject demonstrated *too much* of every professionalism domain. It is likely, then, that the student did not read the anchors for each item and simply gave the highest rating for every subject. The ratings from that student were deleted from the data set. No other anomalies were recognized.

Next, a frequency distribution of the responses across the peer assessments was analyzed. Across the nine questions, 6606 responses were collected (see Table 2). Approximately 88.8 % of the responses were a rating of four, which was the preferred response. The second most identified response was a rating of three (4.8%), followed by a rating of 5 (3.3%), a rating of "not observed" (1.6%), a rating of 2 (.6%), a rating of 6 (.5%), a rating of 1 (.2%), and a rating of 7 (.2%). Scanning the frequency distribution for each question suggests that there was little variability, with approximately 11.2% of the responses deviating from the preferred response (a rating of 4).

Table 2
Frequency Distribution of Responses across the 9 Peer Professionalism Items

Question	Scale Responses								
	0	1	2	3	4	5	6	7	
Honesty	13	0	1	15	606	30	7	2	
Accountability	5	3	8	40	591	21	3	3	
Responsibility	5	2	3	40	608	16	0	0	
Respectful	6	3	3	24	607	26	4	1	
Compassion	7	2	7	37	601	20	0	0	
Maturity	4	0	5	39	601	24	1	0	
Communication	5	0	9	55	563	29	9	4	
Confidentiality	45	0	1	34	593	25	5	5	
Self-directed Learning	10	0	2	34	593	25	5	5	
Total N	100	10	39	291	5381	201	29	15	
Total (%)	(1.6)	(.2)	(.6)	(4.8)	(88.7)	(3.3)	(.5)	(.2)	

The direction of the responses across the peer assessment questions was also analyzed. This was an important analysis because the preferred response is a four, which is the middle of the scale. The semantics of the scale, suggest that students exhibit either too much (5, 6, or 7) or too little (1, 2, or 3) of a particular behavior. For example, if a particular student received peer responses of 2 and a 6 on the accountability question, then that would suggest that peer evaluators disagreed on the direction of the behavior.

For each of the nine questions, 111 students were evaluated by their peers. The number of peer evaluators for each student ranged between 5 and 7 students, resulting in 681 student-peer pairs for each question. Ratings of “0” (not observed) were identified as missing data. There were 104 incidents of students receiving at least 2 ratings that deviated from 4, the preferred response. Twenty of the 104 incidents included responses that suggest rater disagreement on the direction of the scale. For example, one particular student received four 4’s, one 2 and one 6 on the accountability question. Therefore, 4 raters’ responded, with a “preferred” response, 2 raters observed that the student exhibited *too little*, and 1 rater observed that the student demonstrated *too much* of the domain of accountability, which was qualified in the anchors. Approximately 19% of the 104 incidents of at least two ratings that deviated from 4 were identified across all 9 questions, indicating that the majority of students received responses in a singular direction (i.e., one side or the other of a rating of 4). Overall, then, if at least two ratings deviated from the preferred response of 4, then the peer evaluators agreed on the direction (too much or too little) of the professional behavior approximately 80% of the time.

Internal Consistency of the Peer Evaluation - Next, a Cronbach’s alpha was calculated, which addresses the internal consistency of the evaluation. The data set includes responses from multiple raters, which depend on particular subjects. In order to minimize the variance attributed to disagreement between raters, two random samples of the total data set were created. Specifically,

one peer evaluation was randomly selected for each student, totaling 111 peer evaluations. After identifying 0’s (not observed) as missing data, the calculated Cronbach’s alpha was .82. An additional random sample was selected, revealing a comparable alpha of .76. These results suggest that the internal reliability of the responses were moderately consistent.

Generalizability Study - The final analysis targeted inter-rater reliability. One way to investigate inter-rater reliability is to conduct a generalizability study.¹¹ The purpose of a generalizability study is to estimate sources of variance in the absolute mean scores. In this case, raters (r) are nested within subjects, commonly referred to as persons (p). That is, raters only evaluate persons who are in their particular PBL group. Ideally, most of the variation is explained by persons or the “object” of the measurement. The percentage of variance that is explained by persons is analogous to “true score” in classical test theory. The more the variance in responses is explained by persons, the better the reliability estimate. Explaining the technical aspects of a generalizability study is beyond the scope of this paper. Readers may refer to Brennan¹² and Kreiter¹³ to learn more about the types of generalizability studies.

In order to conduct the (r:p) design, an absolute mean score across all nine items was calculated for each rater evaluation of a student. The absolute mean score using this particular scale requires the calculation of a distance score. The ideal is represented by a value of 4. A deviation from this ideal suggests *too much* or *too little* demonstration of a particular domain (e.g., accountability). The description for “too much” is anchored with a value of 7, and the description for “too little” is anchored with a value of 1. When expressed as distance scores, these scores are expressed as +3 and -3, respectively. Distance scores were calculated and converted to an absolute score for each rater evaluation of a student. Finally, the mean absolute distance scores were calculated across all 9 items (see Table 3).

Between 5 and 8 peers evaluated each student. Using the minimum number of raters available for each

Table 3
Generalizability Study Data from Peer Evaluations for 111 Students
in a Problem-Based Learning Course

	Rating			
	Mean Absolute Score	Standard Deviation	Minimum	Maximum
PBL Peer Evaluation	.12	.28	.00	2.29

person, a random sample of five ratings for each student was selected. Using the mean absolute distance score as the rating, the (r:p) generalizability study calculated two variance components: person (.0128) and r:p (.0703). The largest variance component was for rater-nested within person (see Table 4). This result suggests that a student's score depends heavily on a particular rater.

The calculated generalizability coefficient is .48. A decision study (D study) can estimate how implementing

Investigators of this study may also invite other medical schools who might want to utilize the same instrument for evaluation purposes. Once other schools adopt or tailor the instrument, institutional comparisons can be made that might provide additional insights into the instrument's development.

One limitation of the instrument used in this study is the uniformity of responses it generated. Evaluation instruments used by peers and colleagues are especially

Table 4

Variance Component Estimates for (r:p) Design for 5 Peer Professionalism Evaluations of 111 Students in the Fall 2005 Problem-based Learning Course

Source	Variance Component	%
Person (p)	.0128	15
Rater: Persons (r:p)	.0703	85

different designs may affect the generalizability coefficient, which may have implications for optimizing the evaluation. The original coefficient used a design of 5 raters for each person. A D study conducted with 13 raters reveals a projected affect on the generalizability coefficient, which would improve from .45 to .70 (see Figure 1). Increasing the number of raters, then, can reduce error and improve the peer evaluation's measurement precision. Overall, these findings are consistent with previous literature. Specifically, the internal consistency of peer evaluations can be high, while inter-rater reliability is moderate.^{14, 15}

Discussion

Given the exploratory nature of this study, recommendations for future research focus on further investigating the measurement characteristics of the peer professionalism assessment. The first aim for future research is to revise the assessment. For example, a relatively high number of "not observed" responses (45) on the confidentiality question suggest that it may not be appropriate for the context of PBL. That is, students may not be able to judge whether students violated confidentiality issues in a problem-based learning case. However, this question may be used for other situations, such learning opportunities in clinical clerkships.

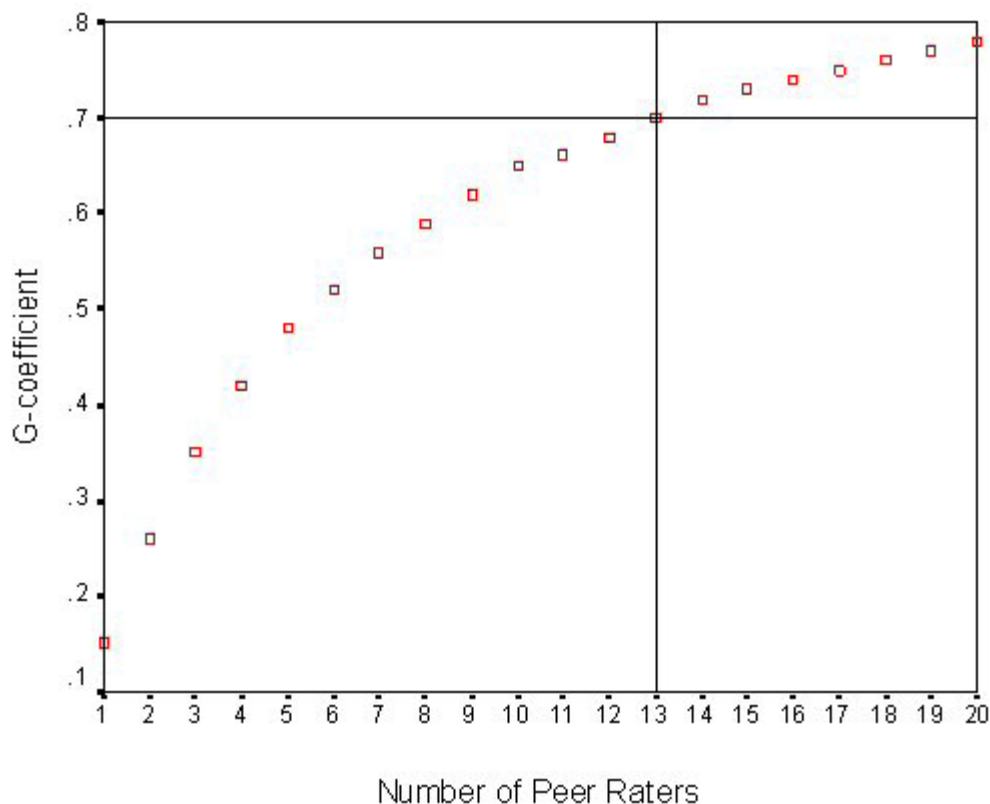
Additional implementations of the peer evaluation are also required. Evaluating professionalism in multiple learning contexts by multiple raters may offer a more precise picture of whether students are demonstrating the professional skills outlined in the professional code.

susceptible to this lack of variance when an ordinal rating scale is provided for each item/characteristic. Raters may tend to avoid "dinging" one another, resulting in repetitively high scores over most items. In addition, there may be a restriction of range relating to the variability of student professionalism. Overall, then, the construct of professionalism has measurement limitations that present challenges to distinguish students' demonstration of professional behavior.

There are a few caveats that should be considered in future implementations of the peer assessment of professionalism. First, the data should be examined to identify whether any students did not understand the bipolar scale. While most students seemed to understand that the preferred response was a "4", scanning the data to identify possible erroneous responses should be conducted before statistical analyses are conducted.

Second, the variance components suggest that the student's score depends heavily the rater. The reliability of a single rating of a student's professional behavior is not sufficient. However, using multiple raters (e.g., 13 or more) for each student does provide a fairly reliable estimate of student's demonstration of professional behaviors. Implementing the peer assessment multiple times may also serve as a training tool. Over time, students can improve their ability to appraise skills and offer constructive feedback. Research has also observed that students value peer feedback, using the results to consider whether particular professionalism attributes are being demonstrated.¹⁶

Figure 1
Generalizability Coefficients of Peer Assessment When Varying the Number of Raters



Overall, these results suggest that this peer assessment is practical. Students are willing to complete the peer evaluation, as evidenced by the 91% compliance rate. This finding is consistent with Arnold's¹⁷ research, which found that students are willing to complete anonymous evaluations of their peers' professionalism. In addition, the results support the conceptualization of professionalism as a complex construct. One evaluation of students' professionalism is not adequate. Implementing the peer assessment multiple times and across a variety of learning contexts lends students opportunities to make formative changes to meet expectations of professional behavior. Coupling this evaluation with other professionalism assessments will help reveal a more complete and distinct picture of students' professionalism.

References

1. Swick HM. Toward a normative definition of medical professionalism. *Academic Medicine*. 2000 Jun;75(6):612-6.
2. Canadian Medical Education Directions for

Specialists (CanMEDS). Extract from the CanMEDS 2000 project societal needs working group report. *Medical Teacher*. 2000;22(6):549-54.

3. Harden RM, Crosby JR, Davis MH, Friedman M. From competency to Meta-competency: a model for the specification of learning outcomes. *Medical Teacher*. 1999;21(6):546-52.
4. Accreditation Council for Graduate Medical Education. c2006. Accreditation Council for Graduate Medical Education Outcomes Project. Available from: <http://www.acgme.org/outcome/>.
5. Association of American Medical Colleges. Report 1: Learning Objectives for Medical Student Education. Guidelines for Medical Schools. 1998.
6. Arnold L. Assessing professional behavior: yesterday, today, and tomorrow. *Academic Medicine*. 2002;77(6):502-15.

7. Lazarus CJ, Chauvin SW, Rodenhauser P, Whitlock R. The program for professional values and ethics in medical education. *Teaching and Learning in Medicine*. 2000;12(4):208-11.
8. Bryan RE, Krych AJ, Carmichael SW, Viggiano TR, Pawlina W. Assessing professionalism in early medical education: experience with peer evaluation and self-evaluation in the gross anatomy course. *Annals of the Academy of Medicine, Singapore*. 2005;34(8):486-91.
9. Devellis RF. *Scale development: Theory and applications*. London: Sage; 2003.
10. Huba ME, Freed JE. *Learner-centered assessment in college campuses: shifting the focus from teaching to learning*. Needham Heights: Allyn & Bacon; 2000.
11. Goodwin LD. Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*. 2001;5(1):13-34.
12. Brennan RL. *Elements of Generalizability Theory*. Iowa City: American College Testing; 1992.
13. Kreiter CD, Ferguson K, Lee W, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. *Academic Medicine*. 1998;73(12):1294-98.
14. Arnold L, Willoughby TL, Calkins EV, Gammon L, Eberhart G. Use of peer evaluation in the assessment of medical students. *Journal of Academic Medicine*. 1981;56:35-42.
15. Panszi S, Gruppen L, Grum C, Stern DT. What do peers know about professionalism? *Proceedings of the Research in Medical Education Conference*. Group on Educational Affairs. AAMC Annual Meeting; 2000; Chicago, ILL, 2000.
16. Dannefer EF, Henson LC, Bierer SB, Grady-Weliky TA, Meldrum S, Nofziger AC, Barclay C, Epstein RM. Peer assessment of professional competence. *Medical Education*. 2005;39:713-22.
17. Arnold L, Shue CK, Kritt B, Ginsburg S, and Stern DT. Medical students' views on peer assessment in professionalism. *Journal of General Internal Medicine*. 2005 Sep;20(9):819-24.

Correspondence

Scott Cottrell, Ed.D.
West Virginia University
Department of Community Medicine
PO Box 9007
Morgantown, WV 26506

Telephone: 304.293.0410
Fax: 304.293.1814
Email: scottrell@hsc.wvu.edu