Relationship Between Peer Assessment During Medical School, Dean's Letter Rankings, and Ratings by Internship Directors

Stephen J. Lurie, MD, PhD, David R. Lambert, MD, Anne C. Nofziger, MD, Ronald M. Epstein, MD, and Tana A. Grady-Weliky, MD

Office of Educational Evaluation and Research, University of Rochester School of Medicine and Dentistry, 601 Elmwood Ave, Box 601, Rochester, NY 14624, USA.

BACKGROUND: It is not known to what extent the dean's letter (medical student performance evaluation [MSPE]) reflects peer-assessed work habits (WH) skills and/or interpersonal attributes (IA) of students.

OBJECTIVE: To compare peer ratings of WH and IA of second- and third-year medical students with later MSPE rankings and ratings by internship program directors.

DESIGN AND PARTICIPANTS: Participants were 281 medical students from the classes of 2004, 2005, and 2006 at a private medical school in the northeastern United States, who had participated in peer assessment exercises in the second and third years of medical school. For students from the class of 2004, we also compared peer assessment data against later evaluations obtained from internship program directors.

RESULTS: Peer-assessed WH were predictive of later MSPE groups in both the second ($F = 44.90$, $P < .001$) and third years ($F = 29.54$, $P < .001$) of medical school. Interpersonal attributes were not related to MSPE rankings in either year. MSPE rankings for a majority of students were predictable from peer-assessed WH scores. Internship directors' ratings were significantly related to second- and third-year peer-assessed WH scores ($r = .32$ [$P = .15$] and $r = .43$ [$P = .004$]), respectively, but not to peer-assessed IA.

CONCLUSIONS: Peer assessment of WH, as early as the second year of medical school, can predict later MSPE rankings and internship performance. Although peer-assessed IA can be measured reliably, they are unrelated to either outcome.

KEY WORDS: assessment; professionalism; undergraduate medical education.
DOI: 10.1007/s11606-007-0117-4

This paper was presented at the Northeast Group on Educational Affairs (Philadelphia, Pa, March 3, 2006) and the Ottawa Conference on Medical Education (New York, NY, May 22, 2006).
Received March 6, 2006
Revised May 12, 2006
Accepted August 29, 2006
Published online January 11, 2007

INTRODUCTION

The dean's letter or "medical student performance evaluation" (MSPE),[1] typically provides a summary of students' performance in clinical settings during medical school. The Association of American Medical Colleges has stressed that the MSPE should be a letter of accurate evaluation rather than of recommendation, and thus that the MSPE should include some form of describing students in comparison with their peers.[1] Despite the fact that dean's letters have repeatedly been found to be variable in terms of their quality[2–7] and accuracy,[8] program directors (PDs) have been found to prefer ranking systems that provide some detail about students' performance relative to their classmates.[9]

We have previously reported that MSPE rankings of medical school graduates are closely related to PDs' later evaluations.[10] It is unclear, however, whether this high-stakes evaluation rewards some attributes more than others. It is possible that attributes relating to work habits (WH) (e.g., organization, efficiency, and knowledge) may be rewarded more than those that relate to more interpersonal ones (e.g., communication, ethics, and empathy). If such attributes do predict later performance, it would be valuable to know how early in training they would become reliably measurable because early identification could allow for timely educational interventions.

We explored these questions by examining the relationships between MSPE rankings (which are determined in the autumn of students' fourth year of medical school) and earlier peer assessment exercises during the second and third years. Similar to other studies of peer assessment,[11] we have found that peer assessment provides a reliable way of measuring the two independent dimensions of WH and interpersonal attributes (IA).[12,13] Because classmates observe one another over larger numbers of occasions and circumstances than do faculty, in theory, classmates should be able to provide valid global assessments of these attributes. MSPE rankings are similarly broadly based, in that they are based on the compilations of a large number of observations over a range of clinical settings. It is unclear, however, whether these two assessment methods produce agreement on relative ratings of students.

Specifically, we attempted to answer two related questions. First, we measured the degree of agreement between the results of peer assessment and later MSPE rankings. Similar results from these different assessment methods would be strong evidence for the validity of peer assessment to evaluate relevant attributes. Such a result would also suggest that these attributes can be measured relatively early in medical training. Second, we examined the degree to which peer

assessment was predictive of later PDs' ratings among a subset of students for whom we had data on both sets of measures. Although there have been several recent calls for increased training in attributes such as communication and ethical behavior, it is unclear to what degree such skills are reflected by the MSPE or valued by PDs. Peer assessment provides a way of distinguishing these elements from the WH dimension.[12,13]

METHOD

Participants

Participants were 281 medical students who graduated in 2004, 2005, or 2006 from the University of Rochester School of Medicine.

Measures

In the final sentence of the MSPE, students were ranked in one of four categories: "outstanding," "excellent," "very good," or "good." Students were assigned to these rankings based upon their grades in required clinical clerkships. In creating the rankings, grades were weighted by the number of weeks of the clerkship and the grade distribution of the entire class in the clerkship. For example, a grade of "honors" in a long clerkship that does not give many honors grades carries more weight than a similar grade in a shorter clerkship that gives many honors grades. The MSPE also provides a guide to interpreting these rankings with approximately 20% in the outstanding group, 25% in the excellent group, and 50–55% in the very good group. Less than 5% are in the good group.

As part of a larger comprehensive assessment program, most students had previously participated in 2 peer evaluation exercises; the first in March of their second year, and the second in June of their third year. At both assessments, students anonymously assessed 6 to 12 classmates. They first completed a standardized rating form[12,13] that assesses the 2 independent factors of professional WH and IA. The six items comprising the IA scale are intermixed with the WH items and include respect; compassion and empathy; seeking to understand others' views; contribution to others' (group's) learning; seeking and responding to feedback; trustworthiness; and honesty in reporting and correcting mistakes. We have previously reported that both scales have a Cronbach alpha of greater than 0.8, that they are only modestly correlated with one another, and that scores in the second year are predictive of scores in the third year, despite the fact that students are generally assessed by different groups of peers in the second and third years.[12] The results of these peer assessments were confidential and not available to the writers of the MSPE at the time these rankings were made.

We sent a 15-item survey to internship program directors approximately 10 months after the students had graduated from medical school. In the survey, PDs were asked to rate the graduates on a number of general clinical, interpersonal, and professional qualities. Factor analysis of this questionnaire was consistent with one-factor solution. Thus, we computed an average of the items to obtain an overall score for PDs' evaluations. At the time of the study these data were only available for members of the class of 2004.

Statistical Analysis

To correct for any systematic differences between classes and year of assessment, we first standardized scores for WH and IA to yield *z*-scores with a mean of 0 and standard deviation of 1.0 for each of the two variables within each of the three classes at each of the 2 years. Our primary analysis involved analysis of variance (ANOVA) to assess the relationship of peer-assessed standardized WH and IA scores to MSPE grouping. Because there were significant differences in peer-assessed WH in the 4 MSPE groups in both the second and third years, we performed a weighted Welch ANOVA adjusting for these differences.

To display MSPE rankings as a function of WH and IA, we performed discriminant function analysis to assess how peer-assessed WH and IA were related to later membership in the four MSPE groups in this sample. Because variances were not homogeneous across the 4 MSPE groups, we used quadratic, rather than linear, discriminant function analysis. Our intent in developing a discriminant function model was not to derive a general prediction rule for predicting MSPE from peer-assessed WH and IA, but rather simply to summarize relationships in our data. This analysis also allowed us to compare how WH and IA performed, both singly and together, in their relationship to MSPE categories.

Relationships between peer assessment and PDs' assessment were assessed by Pearson correlation coefficients. All analyses were performed with SAS version 9.1 (Cary, NC).

RESULTS

Of the 281 graduating students, 41 had not participated in consecutive peer assessment exercises with their classmates. Most of them had taken additional time to pursue another degree or complete a year of research during medical school. This left a total of 240 (85.4%) graduating students who had complete data for the 2 prior peer assessments.

Relationship of MSPE Rankings to Earlier Peer Assessment

The overall multivariate one-way ANOVA for all 4 peer assessment variables (second-year WH and IA and third-year WH and IA) was significant ($F_{12,614} = 9.93$, $P < .001$), which permitted the examination of individual variables. Univariate ANOVA revealed significant differences in peer-assessed WH between the 4 MSPE groups in both the second year ($F_{3,10.8} = 44.90$, $P < .001$) and the third year ($F_{3,9.65} = 29.54$, $P < .001$). Post hoc contrasts using the multiple range test revealed that for both variables the means of the excellent and very good groups were not significantly different than one another. For both variables, the means of these two groups were significantly lower than that of the outstanding group, and significantly higher than that of the good group. Figure 1 portrays the results for the third-year WH scores; results are similar for second year.

By contrast, the four groups did not differ in either the variances or means of their second- or third-year peer-assessed IA. (Regular ANOVA: for second year $F_{3,235} = 1.36$, $P = .26$; for third year $F_{3,235} = .65$, $P = .58$). Results for third-year peer-assessed IA are shown in Figure 2.

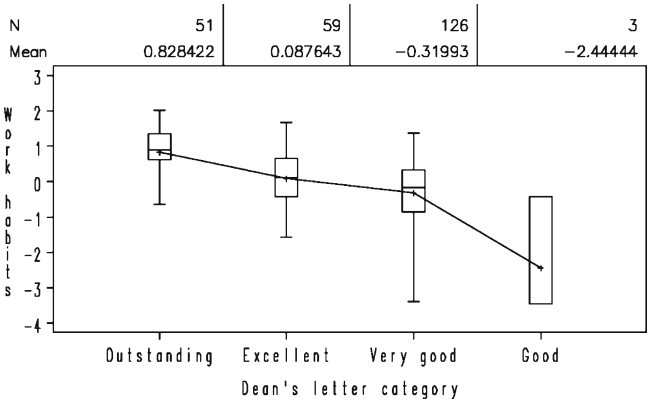

Figure 1. Relationship of peer-assessed WH and MSPE rating for 3 classes of medical students. Within each class, WH scores have been normalized to a mean of 0 and standard deviation of 1.

The discriminant function model (Fig. 3) correctly classified 71% of the students in the outstanding group, 83% of the students in the very good group, and 100% of students in the good group. Of students in the excellent group, 29% were classified as outstanding, 3% as excellent, and 68% as very good. The overall discriminant model had an $F_{12,614} = 9.93$ ($P < .001$, $R^2 = .14$). Quadratic discriminant function analysis using only the 2 IA scores was not statistically significant ($F_{6,468} = .09$, $P = .46$, $R^2 = .01$). The resulting model did not discriminate at all among students and classified most into the most prevalent category of very good. By contrast, a discriminant model using only the WH variables was statistically significant ($F_{6,468} = 18.78$, $P < .001$) and was actually a better fit ($R^2 = .26$) than the full model that had also included IA scores.

Relationship Between Peer Assessment and Later PDs' Reports

There were 43 students from the class of 2004 for whom we received internship directors' ratings (response rate = 44%). Overall ratings were significantly correlated with both second- and third-year WH scores ($r = .32$ [$P = .015$] and $r = .43$ [$P = .004$], respectively). Interpersonal attributes scores were not correlated with later PDs' ratings ($r = .15$ and $-.09$, respectively). Response rates did not differ significantly by MSPE category.

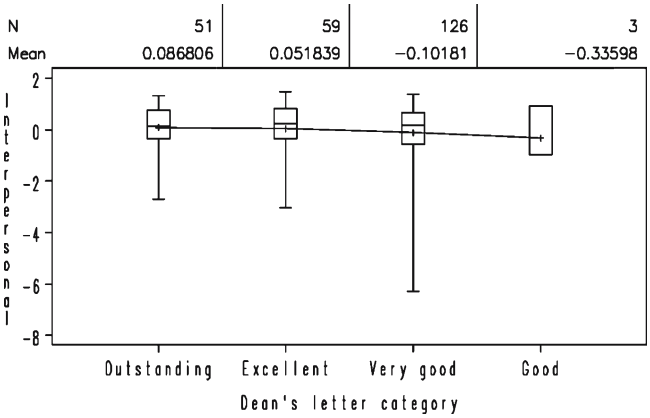

Figure 2. Relationship of peer-assessed IA and MSPE rating for 3 classes of medical students. Within each class, WH scores have been normalized to a mean of 0 and standard deviation of 1.

Actual group	Predicted group: Outstanding	Excellent	Very Good	Good
Outstanding	36 (71%)	3 (6%)	12 (24%)	0 (0%)
Excellent	17 (29%)	2 (3%)	40 (68%)	0 (0%)
Very Good	17 (13%)	5 (4%)	104 (83%)	0 (0%)
Good	0 (0%)	0 (0%)	0 (0%)	3 (100%)

Figure 3. Results of discriminant function analysis for predicting dean's letter rankings from peer assessment data. Numbers in parentheses represent percentages across each row (may not add up to 100 because of rounding).

DISCUSSION

We found that peer assessment among medical students in the second and third years is highly predictive of later MSPE rankings. We have previously reported that MSPE rankings are significantly related to later PDs' assessments.[10] In the current study, we also found that peer assessment was related to later ratings by PDs. These attributes thus appear to be relatively stable from at least the second year of medical school through at least the first year of internship.

In our study these attributes were assessed by four separate groups of raters at four points in time, using three different methodologies. For the 2 peer assessments, scores were generated by different groups of classmates 1 year apart. The MSPE rankings represent a summary of evaluations made by a series of clerkship directors largely during the third year of medical school. Finally, PDs' ratings represent a summary of multiple sets of evaluations during internship, whose details probably vary widely among programs. Nonetheless, these different methodologies all yield remarkably similar rankings, suggesting an overall dimension of global clinical competence that is stable both over time and assessment method.

We note, however, that peer assessment in the third year is a stronger predictor of both MSPE rankings and PD ratings than second-year peer assessment is, although the latter is still a statistically significant predictor. This suggests that peer ratings have greater predictive value when based on observed clinical work rather than classroom work, and/or that students' WH can change during training. This latter possibility suggests that interventions may effect more significant change earlier rather than later in training. Similar observations to ours have been made in the domain of unprofessional behavior. Unprofessional behavior during medical school is a risk factor for later disciplinary action.[14] Early identification can trigger prompt remedial and preventive interventions.[15] We are not aware, however, of any studies that have specifically examined the effects of early intervention on later WH.

Our IA scale has a high internal consistency and is stable when reassessed by different peers using different selection methods. It is only moderately correlated with the peer-assessed WH scale. Thus, we are confident that it represents a stable attribute that is distinct from WH. Nonetheless, we found that peer-assessed IA were unrelated both to MSPE categories and to PD ratings. Such a finding is provocative in light of increasing emphasis in undergraduate medical education on the interrelated areas of communication skills,[16] ethics,[17] professionalism,[18] and psychosocial skills.[19] Further research is needed on the role of early intervention strategies for both poor WH and IA with regard to physicians' overall professional behavior.

There were several students in highest MSPE ranking who had very low levels of peer-assessed IA, as well as many students in the lower MSPE rankings with above-average IA. Perhaps it is not surprising that WH are more valued than interpersonal qualities among clinical clerks and interns, as both deficiencies and excellence in WH are more likely to come to supervisors' attention. Deficiencies in IA may not be as visible to superiors, and may be easier to conceal from teachers than from peers. Traditional high-stakes measures of achievement such as clerkship grades and MSPE rankings do not capture this important dimension. Assessment by peers may provide a measure of these attributes that is otherwise difficult to obtain. Importantly, our peer assessment system is formative, providing written results only to students with the proviso that the report must be discussed with an advisory dean. Students' honesty in rating each other and consequently the robustness of the early predictors of subsequent WH may be compromised if the assessment were to have been conducted differently.

We acknowledge several limitations of our study. First, it was conducted within a single medical school that has a history and a well-established infrastructure for conducting peer assessment. Thus, we believe that our students are relatively well prepared for the task of rating their peers. Institution of peer assessment at other medical schools will need to take students' perspectives into account.[20] Second, our method of assigning MSPE rankings may not represent those of other medical schools. Nonetheless, we point out that we follow AAMC guidelines for preparing these letters, and that these letters are written without any knowledge of how students were rated by their peers. Third, the low response rate among PDs may limit reliability and validity of the correlation that we found between peer assessment and later ratings by PDs. Although the response rate did not differ significantly between the MSPE categories, the possibility of residual confounding remains. Although our rating of IA appears to be reliable, we continue to explore methods of validating it against conceptually similar outcomes. Thus, any findings regarding IA scores remain somewhat difficult to interpret. We believe that this would be an important area for further study.

In summary, our findings suggest that attributes related to clinical WH can be assessed by peers as early as the second year of medical school, and that such assessments should be taken seriously both by students and by their advisors. Such information could provide an "early warning" system for later academic difficulties, which are often more difficult to correct later in training.[21] It is possible that future work will find that such attributes are measurable considerably earlier.

Potential Financial Conflicts of Interest: *None disclosed.*

Corresponding Author: *Stephen J. Lurie Office of Educational Evaluation and Research, University of Rochester School of Medicine and Dentistry, 601 Elmwood Ave, Box 601, Rochester, NY 14624, USA (e-mail: Stephen_Lurie@urmc.rochester.edu).*

REFERENCES

1. American Association of Medical Colleges. A Guide to the Preparation of the Medical Student Performance Evaluation. Available at http://www.aamc.org/members/gsa/mspeguide.pdf. Accessed 6 Jun 2006.
2. **Hunt DD, MacLaren C, Scott C, Marshall SG, Braddock CH, Sarfaty S**. Follow-up study of the characteristics of dean's letters. Acad Med. 2001;76(7):727–33.
3. **Leiden LI, Miller GD**. National survey of writers of dean's letters for residency applications. J Med Educ. 1986;61(12):943–53.
4. **Hunt DD, MacLaren CF, Scott CS, Chu J, Leiden LI**. Characteristics of dean's letters in 1981 and 1992. Acad Med. 1993;68(12):905–11.
5. **Ozuah PO**. Variability in deans' letters. JAMA. 2002;288(9):1061.
6. **Toewe CH 2nd, Golay DR**. Use of class ranking in deans' letters. Acad Med. 1989;64(11):690–1.
7. **Yager J, Strauss GD, Tardiff K**. The quality of deans' letters from medical schools. J Med Educ. 1984;59(6):471–8.
8. **Edmond M, Roberson M, Hasan N**. The dishonest dean's letter: an analysis of 532 dean's letters from 99 U.S. medical schools. Acad Med. 1999;74(9):1033–5.
9. **Provan JL, Cuttress L**. Preferences of program directors for evaluation of candidates for postgraduate training. CMAJ. 1995;153(7):919–23.
10. **Lurie SJ, Lambert DR, Grady-Weliky TA**. Relationship between dean's letter groupings and later evaluations by residency program directors. Manuscript under review.
11. **Norcini JJ**. Peer assessment of competence. Med Educ. 2003;37:539–43.
12. **Lurie SJ, Nofziger A, Meldrum S, Mooney C, Epstein RE**. Temporal and group-related trends in peer assessment amongst medical students. Med Educ. In press.
13. **Dannefer EF, Henson LC, Bierer SB, et al**. Peer assessment of professional competence. Med Educ. 2005;39:713–22.
14. **Papadakis M, Teherani A, Banach MA, et al**. Disciplinary action by medical boards and prior behavior in medical school. N Engl J Med. 2005;353(25):2673–82.
15. **Loeser H, Papadakis M**. Promoting and assessing professionalism in the first two years of medical school. Acad Med. 2000;75(5):509–10.
16. Association of American Medical Colleges. *Contemporary Issues in Medicine: Communication in Medicine*. Washington, DC: Association of American Medical Colleges; 1999. Report 3 of the Medical School Objectives Project. Available at http://www.aamc.org/meded/msop/msop3.pdf. Accessed 6 Jun 2006
17. **DuBois JM, Burkemper J**. Ethics education in U.S. medical schools: a study of syllabi. Acad Med. 2002;77(5):432–7.
18. **Veloski JJ, Fields SK, Boex JR, Blank LL**. Measuring professionalism: a review of studies with instruments reported in the literature between 1982 and 2002. Acad Med. 2005;80(4):366–70.
19. **Kern DE, Branch WT Jr, Jackson JL, et al**. General Internal Medicine Generalist Educational Leadership Group. Teaching the psychosocial aspects of care in the clinical setting: practical recommendations. Acad Med. 2005;80(1):8–20.
20. **Arnold L, Shue CK, Kritt B, Ginsburg S, Stern DT**. Medical students' views on peer assessment of professionalism. J Gen Intern Med. 2005;20(9):819–824.
21. **Dudek NL, Marks MB, Regehr G**. Failure to fail: the perspectives of clinical supervisors. Acad Med. 2005;80(10 suppl):S84–7.