Outcome Measurement in Postgraduate Year One of Graduates from a Medical School with a Pass/Fail Grading System

Kenneth L. Vosti, MD, and Charlotte D. Jacobs, MD

ABSTRACT

Purpose. To measure the performances of first-year residents who had graduated from a medical school with a pass/fail grading system and to compare the preparedness of these graduates with that of their peers.

Method. All 169 graduates of Stanford University School of Medicine's classes of 1993 and 1994 were included in this study. First-year program directors rated the performance of each Stanford graduate in 11 areas, compared the graduate's clinical preparedness with that of his or her peer group, and rated the accuracy of the dean's letter in presenting the graduate's capabilities.

Results. Responses were obtained for 144 of the 169 graduates (85%). The program directors rated the overall clinical competencies of most of the graduates as "superior" (76%) or "good" (22%); they rated very few as "unsatisfactory" (2%). When the Stanford graduates were compared with their peers, their clinical preparedness was judged "outstanding" (33%), "excellent" (44%), and "good" (20%); very few were judged "poor" (3%). Stratification of programs by either hospital or medical specialty did not reveal significant differences in overall clinical competence. Ninety-one percent of the responses reported that the dean's letters had accurately presented the capabilities of the graduates.

Conclusion. Graduates from a medical school with a two-interval, pass/fail system successfully matched with strong, highly-sought-after postgraduate training programs, performed in a satisfactory to superior manner, and compared favorably with their peer group.

Acad. Med. 1999;74:547–549.

The importance of outcome measurements in assessing the effectiveness of educational programs was emphasized recently by Kassebaum.[1] Institutions with pass/fail systems for rating students' performances are especially challenged by such assessments of educational outcome and are confronted with at least two significant limitations: (1) the quantitation of the effectiveness of their educational programs, and (2) the presentation of their students to selection committees of postgraduate training programs. Thus, the absence of grades, honors, and class rankings presents a challenge both to those who recommend students to and to those who select them for postgraduate training programs. Despite these limitations, Stanford University School of Medicine has employed a pass/fail system, without honors or class rankings, for rating our students' performances in both the basic and clinical sciences since 1968.

In the present study, we surveyed postgraduate training program directors to quantify their assessments of the performances of our graduates in their programs, of the clinical preparedness of our graduates as compared with that of their peer group, and of the accuracy of the dean's letter.

METHOD

In this study, we included all 169 graduates from the Stanford University School of Medicine's classes of 1993 and 1994 who sought further clinical training. We asked the directors of the students' first-postgraduate-year (PGY-1) training programs to complete an evaluation form for each of our graduates in their programs. Of the 169 evaluation forms sent, the program directors returned 144 (85%).

The program directors (or their delegates) rated the performances of the graduates during the last quarter of their PGY-1 training in six cognitive skills (history taking, physical examination, performance of procedures, medical knowledge, clinical judgment, and ability to apply knowledge) and four noncognitive skills (interpersonal rela-

Dr. Vosti is professor of medicine and associate dean for student affairs, emeritus, and Dr. Jacobs is professor of medicine and former senior associate dean for medical education and student affairs, both at Stanford University School of Medicine, Stanford, California.

Correspondence and reprint requests should be addressed to Dr. Vosti, Division of Infectious Diseases, S156, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305.

tionships with professionals, patients, and peers, and dependability), and in their overall clinical competence.

The rating scale ranged from 1 to 9, with 1–3 identified as unsatisfactory, 4–6, satisfactory, and 7–9, superior. At the two ends of the scale, characteristics that defined the lowest and highest performance levels were briefly described. The program director based the ratings on observations provided by one or more attending physicians who had worked with the graduate. In addition, the director compared the graduate's clinical preparedness on entry into the program with that of his or her peer group (relatively poor, good, excellent, or outstanding) and responded "yes" or "no" to the question of whether the dean's letter had accurately presented the graduate's capabilities.

Using GraphPad Prism version 2.0 (GraphPad Software Inc., San Diego, CA), we performed both standard column statistical analyses and, to compare the graduates by hospital group and medical specialty, the Kruskal–Wallis test with Dunn's post-multiple pairwise comparison test.

Results

Population characteristics. The 144 graduates for whom we received evaluation forms were a diverse group. They included 56 (39%) women and 88 (61%) men. Eighty-six (60%) of the graduates were white; 33 (23%), Asian/Pacific Islanders; 11 (8%), Hispanic; seven (5%), black; and four (3%), Native Americans. Three (2%) belonged to other ethnic groups. Of the 144 graduates, 106 (74%) were in university programs, and 38 (26%) were in university-affiliated training programs in community hospitals. The graduates attended 52 different graduate programs in 19 different areas of training: internal medicine (27), preliminary internal medicine (19), pediatrics (18), orthopedic surgery (11), family practice (9), transitional (9), emergency medicine (8), general surgery (8), obstetrics and gynecology (7), pathology (7), and nine other areas (21).

Clinical performance. Ratings for seven of the 11 variables ranged from 3 to 9; ratings for the other four variables ranged from 4 to 9. The median rating for each of the cognitive variables and the overall assessment was 7; for each of the noncognitive variables, 8. Mean ratings for cognitive skills ranged from 6.86 ± 1.21 for procedures to 7.21 ± 1.40 for application of knowledge; and, for noncognitive skills, from 7.60 ± 1.40 to 7.76 ± 1.22.

Ratings of overall clinical competence ranged from 3 to 9, with a median of 7 and a mean (± SD) of 7.19 ± 1.28. Only three (2%) of the 143 graduates received an overall assessment of "unsatisfactory" (rating of 3); in contrast, 108 (76%) received "superior" ratings (ratings of 7–9).

Other ratings. When the program directors compared our graduates' clinical preparedness with that of their peer group, 4/143 (3%) were rated "poor"; 29 (20%), "good"; 63 (44%), "excellent"; and 47 (33%), "outstanding". Of the 141 evaluation forms that included a response to the question, 129 (91%) judged the dean's letter to have accurately presented the capabilities of the graduates. Two of the 12 negative responses stated that the dean's letter had underrepresented the capabilities of the graduate.

Hospital groups and medical specialties. We divided the graduates into four hospital groups: our own (Stanford Medical Center), an example of strong, highly-sought-after programs (Harvard Hospitals), all other university hospitals, and community hospitals. Similarly, we grouped the graduates by their medical specialties. Table 1 presents summary statistics for the ratings of the graduates' overall clinical performances and for the comparison of our graduates' clinical preparedness with that of their peers. Although variations in the mean overall ratings existed among the hospital groups, no significant pairwise difference ($p > .05$) was found when analyzed by the nonparametric Kruskal–Wallis test with Dunn's post-multiple pairwise comparison test. Nor did similar analyses of the other categories identify any significant difference.

Discussion

Changes in the grading of medical students' performances have created significant controversy and strong opinions for and against variations in and departures from the traditional methods of grading students.[2–7] A recent survey by the AAMC of 128 medical schools in the United States revealed that only 5% used two grading intervals to grade students in their required clinical clerkships; 22%, three grading intervals; 21%, four grading intervals; and 38%, five grading intervals.[8] A similar distribution was found for the required basic science courses.

Although this controversy has existed for over 30 years, evidence favoring one or another of the various rating systems is limited. After a brief trial, one prestigious medical school dropped its pass/fail grading system (along with major changes it had made to its basic science curriculum) because of its faculty's increasing concern and a decline in its students' scores on Part I of the National Board examinations.[3] We suspect that the curricular changes were more likely the cause than was the pass/fail system, since the school subsequently returned to a pass/fail system for the basic sciences.[9]

Based on a retrospective study of the performances of residents in a general surgery program, Moss and colleagues recommended that program directors preferentially select graduates from schools that grade their students along a more traditional range. Although all of the residents in that study had satisfactory overall performances, the authors found that residents whose dean's letters included "class ranking" received significantly higher ratings in overall

Table 1

Comparison of PGY-1 Ratings Given by Program Directors to Graduates of Stanford University School of Medicine, by Hospital Group and Medical Specialty

	No. of Responses	Rating	
		Median (Range)	Mean ± SD
Overall ratings			
All graduates	143	7 (3–9)	7.19 ± 1.28
Hospital group			
Stanford Medical Center	29	8 (6–9)	7.62 ± 0.98
Harvard Hospitals*	23	8 (6–9)	7.71 ± 0.98
Other university hospitals	53	7 (3–9)	6.81 ± 1.50
Community hospitals	38	7 (3.5–9)	7.07 ± 1.12
Medical specialty programs			
Internal medicine	27	7 (5–8)	6.98 ± 1.04
Preliminary internal medicine	19	7 (5–9)	7.13 ± 1.01
Pediatrics	18	7.75 (3.5–9)	7.50 ± 1.41
General surgery	8	7 (3–9)	6.50 ± 1.77
Surgical subspecialties†	24	7.5 (4–9)	7.38 ± 1.43
Comparison of Stanford graduates' clinical preparedness upon entering PGY-1 program with that of their peers			
All graduates	143	3 (1–4)	3.05 ± 0.80
Hospital group			
Stanford Medical Center	29	3 (2–4)	3.21 ± 0.77
Harvard Hospitals*	23	3 (2–4)	3.17 ± 0.79
Other university hospitals	54	3 (1–4)	2.93 ± 0.87
Community hospitals	37	3 (1–4)	3.04 ± 0.71
Medical specialty programs			
Internal medicine	27	3 (2–4)	2.93 ± 0.68
Preliminary internal medicine	19	3 (2–4)	3.26 ± 0.81
Pediatrics	18	3 (1–4)	3.19 ± 0.89
General surgery	8	3 (2–3.5)	2.69 ± 0.59
Surgical subspecialties†	24	3 (1–4)	3.13 ± 0.90

*Harvard Hospitals: Brigham and Women's Hospital (9), Boston Children's Hospital (7), Massachusetts General Hospital (6), and Deaconess Hospital (1).

† Surgical specialties: orthopedic (11), plastic (5), otolaryngology (3), urology (3), and neurosurgery (2).

PGY-1 performance than did those whose letters contained only "descriptive prose" (mean rating 4.14 ± 0.28 vs. 3.88 ± 0.31, $p < .001$).[5] Although that study is cited as a condemnation of the pass/fail system, the authors labeled all residents whose letters did not contain a class ranking as "pass/fail"; almost certainly, that group was not composed solely of students from schools using a grading interval of two. Additional concerns arise when this study is used to reflect the use of the pass/fail system in individual medical schools. A later survey of 760 residency program directors from diverse specialties revealed that 73% did not give preference to candidates from either traditionally graded or pass/fail schools.[6]

In the present study, we found that graduates from a single medical school with a two-interval, pass/fail grading system: (1) can successfully match with strong, highly-sought-after postgraduate training programs, (2) usually perform in a satisfactory-to-superior manner, and (3) compare favorably with their peer group in terms of their clinical preparedness. We recognize the complexity and potential bias that may influence such analyses; however, we found no significant difference in the results when the graduates were stratified by hospital groupings or selected medical specialties.

We interpret these external outcome measurements of our graduates' performances in PGY-1 as strong support for the effectiveness of our educational program, for the process used in counseling and assisting students in matching successfully with appropriate postgraduate programs, and for our continued use of a pass/fail grading system. Finally, we believe that periodic surveys of outcome measurements similar to ours would provide an important quality-control program not only for institutions with pass/fail grading systems but also for those with more numerous grading intervals.

References

1. Kassebaum DG. The measurements of outcome in the assessments of educational program effectiveness. Acad Med. 1990;65:293–6.
2. Bender RM. Attitudes toward grading systems used in medical education. J Med Educ. 1969; 44:1076–81.
3. Goldhaber SZ. Medical editorial: Harvard reverts to tradition. Science. 1973;181:1027–32.
4. Stimmel B. The use of pass/fail grades to assess academic achievement and house staff performance. J Med Educ. 1975;50:657–61.
5. Moss TJ, Deland EC, Maloney JV. Selection of medical students for graduate training: pass/fail versus grades. N Engl J Med. 1978;299:25–7.
6. Tardiff K. The effect of pass-fail on the selection and performance of residents. J Med Educ. 1980;55:656–61.
7. Weiss ST, Rosa RM, Jofe T, Munoz B. A prospective evaluation of performance during the first year of the medical residency. J Med Educ. 1984;59:967–8.
8. 1996–1997 AAMC Curriculum Directory. Washington, DC: Association of American Medical Colleges, 1997: 8–10, table 1.
9. 1991–1992 AAMC Curriculum Directory. Washington, DC: Association of American Medical Colleges, 1991:113.