REPRODUCIBILITY OF ORAL EXAM GRADES AND CORRELATION WITH OTHER MEASURES OF PERFORMANCE ON THREE REQUIRED THIRD-YEAR CLERKSHIPS

L. PETER SCHWIEBERT
ALAN B. DAVIS
M. ALEX JACOCKS
University of Oklahoma

The oral examination is one of the traditional measures of student performance during clinical clerkships. Other studies have compared oral exams, written exams, and clinical performance, finding an unequal correlation among them and poor reproducibility of scores among examiners. This study of a third-year class on three required clerkships found a stronger correlation between oral exam performance and cumulative grade point average (GPA) than had previously been reported between oral exams and written or clinical grades and also found high reproducibility across clerkships, both overall and within class quartiles. These findings argue for wider use of the oral exam as an evaluation instrument on clinical clerkships.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 15 No. 2, June 1992 221-230

The oral exam is not as commonly used as other measures of student performance, such as written exams and the clinical evaluation (Beckmann, Barzansky, Eden, Ling, & Waxman, 1989; Magarian & Mazur, 1990). One problem cited with the oral exam and leading to less widespread use is lack of reproducibility (Bull, 1959; Colton & Peterson, 1967; Kelley, Matthews, & Schumacher, 1971; McCormick, 1981), where reproducibility refers to getting similar results when two or more examiners evaluate an examinee. However, the oral exam continues to be used and recommended, because it is said to measure aspects of student performance that are not covered in clinical and written evaluations (Beckmann et al., 1989; Doyle, 1980; Halio, 1963; Levine & McGuire, 1970; Magarian & Mazur, 1990; Stebbins, 1951; Yang & Laube, 1953; Zelenock et al., 1985). The oral exam is felt to assess aspects of personality, including alertness and stress tolerance (Doyle, 1980), ability to respond to changes in situation (Levine & McGuire, 1970), mental agility (Halio, 1963), and poise (Stebbins, 1951). Other authors also claim the oral exam measures problem-solving skills—ability to "think on one's feet" (Zelenock et al., 1985), to organize ideas and grasp a point quickly (Stebbins, 1951), and to process information (Yang & Laube, 1953). In addition, the oral exam may be helpful to students who have difficulty expressing themselves in writing (Moore-Rinvolucri & Nixon, 1952-1953) and provides valuable faculty-student interaction (Zelenock et al., 1985).

At the University of Oklahoma Health Sciences Center, three required third-year clerkships culminate in the evaluation of students with an oral exam; these three clerkships are surgery, family medicine, and obstetrics. The processes for each clerkship's oral exam are compared in Table 1. The three exams differ in degree of structure, with the family medicine exam being the most structured—suggesting essential responses, for example; surgery and obstetrics are less structured—suggesting topics to cover or providing case scenarios, respectively.

Because there had been no effort to standardize the way these exams were administered and because of the problems with reproducibility found in other studies (Bull, 1959; Colton & Peterson, 1967; Kelley et al., 1971; McCormick, 1981), it was hypothesized that the current study would show poor correlation of oral exam scores across the three

TABLE 1
A Comparison of Oral Exam Operation on Three Clerkships

	Surgery	*Obstetrics*	*Family Medicine*
Examiners	5 or 6 teams of 2 examiners (1 faculty plus 1 resident of 2 faculty/team).	1 team (2 faculty plus 1 resident) examines all students.	2 exam sessions; 3 examiners/session (2 faculty & 1 resident); ≥2 of these 3 examiners participate in both sessions.
Information students are supplied to prepare for the exam	List of 5 basic surgical topics is supplied at the start of month. At the start of exam, students learn which 2 topics of these are on their exam. Third topic selected by examiners – may or may not be on 5-topic list.	75 case scenarios supplied at start of month. Students are given their 3 cases from this list 15 minutes before exam.	27 core topics and learning objectives supplied at start of month. 20 minutes before exam, students are given their 3 cases and a question for each case.
Exam structure	Non-directive – examiners may ask different questions of each examinee.	Questions for each case are defined, but appropriate responses are not.	Questions and essential appropriate responses are defined; examiners are encouraged to be open to nonlisted responses.
Exam Duration	30 min	15 min	30 min
Grading process	Each examiner assigns the student a letter grade; the student's raw exam grade is the mean of examiners' grades. Course director may adjust this grade based on student's overall performance.	Each examiner assigns a number between 1 and 4 for student's performance on each case. The mean score for 3 examiners is calculated for each case & from this a mean for the 3 cases is calculated.	100 points are allocated for 3 cases and a specific number of points is allowed for each question.Each student's raw score (3 examiners' total points + 300) is adjusted based on difference in difficulty between 2 days' exams and examiner consensus on letter grade earned by examinee with top score.

clerkships, confirming the findings of other investigators. In addition, this study seeks answers to the following questions, in the hope this information will prove helpful to those planning and implementing oral exams:

1. To what extent does student performance on the oral exam correlate with overall performance?
2. What impact does structure have on the validity of the oral exam?

METHODS

Oral exam grades for each student in the 1989-1990 junior medical school class (89 students) were compared across the surgery, family medicine, and obstetrics clerkships and were also compared with students' cumulative medical school grade point average (GPA) at the end of third-year coursework. An advantage to comparing grades across clerkships instead of within one clerkship is elimination of the spuriously high agreement that can occur when examiners have the opportunity to consult with one another prior to assigning a score (McGuire, 1966).

All three clerkships assign students a letter grade; however, one (family medicine) assigns a score on a 100-point scale, one (obstetrics) assigns a score on a 4-point scale (e.g., 3.41), and one (surgery) assigns a letter grade only. Because comparison could not be more precise than the broadest score of the three, the obstetrics and family medicine grades were converted to letter grades and comparisons in this study were based on a 5-point integer scale, where F = 0 and A = 4.

Agreement was defined as two or more of the three clerkships assigning the same letter grade to a given student. *Disagreement* was defined as none of the three assigning a student the same letter grade. Grade comparisons across the clerkships were made for each student by indicating whether two or more agreed, all three agreed, or none agreed. These comparisons were by quartile as well as for the class as a whole. To assess the significance of these comparisons in the context of other measures of student performance, the same comparisons were made for each student's written and clinical grades.

TABLE 2
Agreement on Oral Exam Grades Among Three Clerkships for a Class of Third-Year Medical Students

	≥ 2 Clerkships Agree	3 Clerkships Agree	None Agree
Overall	72/89	17/89	17/89
(89 Students)	81%	19%	19%
First Quartile	22/25	10/25	3/25
(25 Students)	88%	40%	12%
Second Quartile	16/20	2/20	4/20
(20 Students)	80%	10%	20%
Third Quartile	18/20	5/20	2/20
(20 Students)	90%	40%	10%
Fourth Quartile	16/24	0/24	8/24
(24 Students)	67%	0%	33%

For correlation with students' overall performance, GPA was selected, because it is the only standard measure of academic performance against which student performance in all three clerkships can be compared. GPA for each student was recorded to hundredths of a point, and oral exam scores were recorded as integers (A = 4, B = 3, etc.). For mean GPA, the sum of all students' GPAs was divided by the number of students; likewise the mean oral exam grade for each clerkship was calculated by dividing the sum of all oral exam grades on the clerkship by the number of students. This calculation was also carried out for each quartile of the class. Correlations among mean GPAs and oral exam grades were compared, using a Pearson correlation coefficient.

RESULTS

Two or more of the clerkships agreed on the oral exam grade for 81% of the students; therefore in 19% of cases, none of three clerkships agreed (Table 2). This compared to 74% agreement on the written exam and 96% agreement on clinical grades. On the oral exam, 22% of grades on the three clerkships were less than B; this figure was 37% and 7% for written and clinical evaluations, respectively.

TABLE 3
Grade Distribution for a Third-Year Class – Grade Point Average (GPA) and Oral Exam Scores on Three Clerkships

	Number of Students Receiving:				
	A	*B*	*C*	*D*	*F*
GPA (89 Students)	25 (28%)	60 (67%)	4 (5%)	0	0
Family Medicine Oral Exam (89 Students)	22 (25%)	26 (29%)	32 (36%)	7 (8%)	2 (2%)
Obstetrics Oral Exam (87 Students)	50 (57%)	30 (34%)	7 (9%)	0	0
Surgery Oral Exam (88 Students)	23 (26%)	54 (61%)	11 (13%)	0	0

The proportion of students with the same oral exam grade on two or more clerkships was higher for the top three quartiles of the class than the fourth quartile (Table 2). Comparing proportional agreement for the top three quartiles with the fourth quartile resulted in a significant p value ($p = .038$), indicating evaluators had more difficulty agreeing on the performance of students toward the bottom of the class, compared to students higher in the class.

Grade distribution for oral exams and GPA for the 89 students is shown in Table 3. Ninety-five percent of the students had an A or B GPA, while 54%, 91%, and 87% of the students received an A or B on family medicine, obstetrics, and surgery oral exams, respectively.

Mean oral exam scores are compared with mean GPA in Table 4. Mean oral exam scores in all three clerkships increase with increasing class quartile. On the less structured oral exams, mean oral exam grades are not significantly different from overall GPA and from second and third quartile GPAs (surgery) and first quartile GPAs (obstetrics). Mean oral exam grades on the more structured family medicine clerkship were significantly different overall and in each quartile from mean GPA.

Oral exam grades on all three clerkships had a statistically significant correlation with GPA. The most structured exam (family medicine) had the strongest overall correlation with GPA, .576 ($p < .0001$).

TABLE 4
Comparison Between Mean Oral Exam Scores and Grade Point Average (GPA)

	Oral Exam Scores			
	Family Medicine	*Obstetrics*	*Surgery*	*GPA*
Overall (89 Students)	2.70	3.49	3.11*	3.14*
First Quartile (25 Students)	3.48	3.81*	3.40	3.79*
Second Quartile (20 Students)	2.95	3.75	3.25*	3.28*
Third Quartile (20 Students)	2.48	3.32	3.09*	3.02*
Fourth Quartile (24 Students)	2.04	3.17	2.80	2.57

*Nonsignificant difference between oral exam scores and GPA.

Obstetrics was next with a correlation coefficient of .421 ($p < .0001$), and the correlation coefficient for surgery was .351 ($p < .001$). All of these correlations are generally stronger than those reported between oral and written exams and between the oral exam and clinical evaluation (Bull, 1959; Colton & Peterson, 1967; DiNio, Holmes, Pierleoni, & Greenberger, 1975; Evans, Ingersoll, & Smith, 1966; Ginsburg, 1985; Littlefield, Harrington, & Garman, 1977; O'Donoghue & Wergin, 1978; Zelenock et al., 1985).

DISCUSSION

This study found oral exam grades were more reproducible than written exam grades but less reproducible than clinical grades across three clerkships. Unlike others, this study reported agreement of oral exam grades within class quartile and found this was also high, except in the fourth quartile. It is unclear why fourth quartile grades are less reproducible; perhaps this is due to some evaluators' reluctance to assign lower grades or to those students performing less consistently than their upper quartile peers.

The much higher level of agreement for clinical than for oral or written grades is due to an essentially dichotomous grade distribution (As and Bs) among clinical grades; the reduction in range for the clinical grades increases the likelihood at least two clerkships will award a student the same grade.

The degree of reproducibility of oral exam grades is surprising, both in view of the findings of most other authors and the lack of control of methodologies across the three clerkships' oral exams. This reproducibility across a class of 89 students indicates the oral exam must measure certain common attributes among students and argues for wider use of the oral exam as an evaluation instrument.

This study also compared oral exam and overall performance by correlating oral exams with GPA. Among a class of 89 third-year medical students, a significant correlation was found between mean overall GPA and oral exam grades on all three clerkships; this correlation is stronger than previously reported correlations between oral and written exams and oral and clinical evaluations (Bull, 1959; Colton & Peterson, 1967; DiNio et al., 1975; Evans et al., 1966; Ginsburg, 1985; Littlefield et al., 1977; O'Donoghue & Wergin, 1978; Stebbins, 1951; Zelenock et al., 1985) and occurred both with relatively structured and unstructured oral exams. This indicates the oral exam is a valid measure of academic performance and also argues for wider use of the oral exam as an evaluation tool.

The obvious explanation for the close resemblance between mean surgery and obstetrics oral exam grades and GPA is the strong skew of all three toward As and Bs (Table 3). It is more difficult to explain why the two less structured exams have this skew, whereas the more structured family medicine exam shows a preponderance of midrange grades (64% Bs and Cs). A possible explanation is that the more structured exam, by providing examiners specific responses to listen for, may more critically evaluate the student's fund of knowledge and clinical reasoning than a less structured exam, where the student's enthusiasm and motivation may more strongly influence grading—the so-called "halo effect" well-documented in clinical grading (Marienfeld & Reid, 1980, 1984; Quarrick & Sloop, 1972). On the other hand, a structured exam risks unfairly penalizing students by expecting too-extensive responses from them, given the time limits and pressures

inherent in the exam process. Based on the first year's experience with the family medicine oral exam, course planners are reevaluating grade distribution and the amount of content students are expected to cover during a 30-minute exam.

The occurrence of the strongest oral exam-GPA correlation coefficient on the most structured exam argues for a more structured approach to oral exam design, if correlation with overall performance is a goal. Indeed, other authors (Evans et al., 1966; Yang & Laube, 1953) have found a more structured approach improves interrater reliability, but none have examined the impact of structure on correlation with overall student performance. Further studies comparing more and less structured oral exams with global performance are needed to validate this finding.

REFERENCES

Beckmann, C.R.B., Barzansky, B. M., Eden, R. D., Ling, F. W., & Waxman, B. (1989). Student evaluation in obstetrics and gynecology clerkships in the United States and Canada, 1985. *Journal of Reproductive Medicine, 34*, 349-352.

Bull, G. M. (1959). Examinations. *Journal of Medical Education, 34*, 1154-1158.

Colton, T., & Peterson, O. L. (1967). An assay of medical students abilities by oral examination. *Journal of Medical Education, 42*, 1005-1014.

DiNio, J. N., Holmes, F. F., Pierleoni, R. G., & Greenberger, N. J. (1975). *Evaluation of internal medicine clerkship students*. Paper presented at the 14th annual conference on Research in Medical Education, Washington, DC.

Doyle, M. (1980). *Oral examinations—the current state of the art: Where do we go from here?* Manuscript submitted to the R. S. McLaughlin Examination and Research Center.

Evans, L. R., Ingersoll, R. W., & Smith, E. J. (1966). The reliability, validity, and taxonomic structure of the oral examination. *Journal of Medical Education, 41*, 651-657.

Ginsburg, A. D. (1985). Comparison of intraining evaluation with tests of clinical ability in medical students. *Journal of Medical Education, 60*, 29-36.

Halio, J. L. (1963). Ph.D.'s and the oral examination. *Journal of Higher Education, 34*, 148-152.

Kelley, P. R., Matthews, J. H., & Schumacher, C. F. (1971). Analysis of the oral examination of the American Board of Anesthesiology. *Journal of Medical Education, 46*, 982-988.

Levine, H. G., & McGuire, C. H. (1970). The validity and reliability of oral examinations in assessing cognitive skills in medicine. *Journal of Educational Measurement, 7*, 63-73.

Littlefield, J. H., Harrington, J. T., & Garman, R. E. (1977). *Use of an oral examination in an internal medicine clerkship*. Paper presented at the 16th annual conference on Research in Medical Education, Washington, DC.

Magarian, G. J., & Mazur, D. J. (1990). Evaluation of students in medicine clerkships. *Academic Medicine, 65*, 341-345.

Marienfeld, R. D., & Reid, J. C. (1980). Subjective vs. objective evaluation of clinical clerks (letter). *New England Journal of Medicine, 302*, 1036-1037.

Marienfeld, R. D., & Reid, J. C. (1984). Six-year documentation of the easy grader in the medical clerkship setting. *Journal of Medical Education, 59*, 589-591.

McCormick, W. O. (1981). A practice oral examination rating scale – inter-observer reliability. *Canadian Journal of Psychiatry, 26*, 236-239.

McGuire, C. H. (1966). The oral examination as a measure of professional competence. *Journal of Medical Education, 41*, 267-274.

Moore-Rinvolucri, M. J., & Nixon, B. M. (1952-1953). How to improve oral questioning. *Peabody Journal of Education, 30*, 209-217.

O'Donoghue, W. J., & Wergin, J. F. (1978). Evaluation of medical students during a clinical clerkship in internal medicine. *Journal of Medical Education, 53*, 55-58.

Quarrick, E. A., & Sloop, E. W. (1972). A method for identifying the criteria of good performance in a medical clerkship program. *Journal of Medical Education, 47*, 188-197.

Stebbins, E. L. (1951). Panel discussion: Objective or multiple choice type of examination, American Board of Preventive Medicine and Public Health. In *Proceedings of the Annual Congress on Medical Education and Licensure* (pp. 53-57). Chicago: American Medical Association.

Yang, J. C., & Laube, D. W. (1953). Improvement of reliability of an oral examination by a structured evaluation instrument. *Journal of Medical Education, 58*, 864-872.

Zelenock, G. B., Calhoun, J. G., Hockman, E. M., Youmans, L. C., Erlandson, E. E., Davis, W. K., & Turcotte, J. G. (1985). Oral examinations: Actual and perceived contributions to surgery clerkship performance. *Surgery, 97*, 737-743.