Prospective Implementation of Correction for Guessing in Oral and Maxillofacial Pathology Multiple-Choice Examinations: Did Student Performance Improve?

Thomas J. Prihoda, Ph.D.; R. Neal Pinckard, Ph.D.; C. Alex McMahan, Ph.D.; John H. Littlefield, Ph.D.; Anne Cale Jones, D.D.S.

Abstract: A standard correction for random guessing on multiple-choice examinations was implemented prospectively in an Oral and Maxillofacial Pathology course for second-year dental students. The correction was a weighted scoring formula for points awarded for correct answers, incorrect answers, and unanswered questions such that the expected gain in the multiple-choice examination score due to random guessing was zero. An equally weighted combination of four examinations using equal numbers of short-answer questions and multiple-choice questions was used for student evaluation. Scores on both types of examinations, after implementation of the correction for guessing on the multiple-choice component (academic year 2005–06), were compared with the previous year (academic year 2004–05) when correction for guessing was not used for student evaluation but was investigated retrospectively. Academically, the two classes were comparable as indicated by the grade distributions in a General Pathology course taken immediately prior to the Oral and Maxillofacial Pathology course. Agreement between scores on short-answer examinations and multiple-choice examinations was improved in the 2005–06 class compared with the 2004–05 class. Importantly, the test score means were higher on both the short-answer and multiple-choice examinations in the Oral and Maxillofacial Pathology course, and the standard deviations were significantly smaller in 2005–06 compared to 2004–05; these differences reflected an upward shift in the lower part of the grade distributions to higher grades in 2005–06. Furthermore, when students were classified by their grade in the General Pathology course, students receiving a C (numerical grade of 70–79 percent) in General Pathology had significantly improved performance in the Oral and Maxillofacial Pathology course in 2005–06, relative to 2004–05, on both short-answer and multiple-choice examinations representing an aptitude-treatment interaction. We interpret this improved performance as a response to a higher expectation imposed on the 2005–06 students by the prospective implementation of correction for guessing.

Dr. Prihoda is Associate Professor, Department of Pathology; Dr. Pinckard is Professor, Department of Pathology; Dr. McMahan is Professor, Department of Pathology; Dr. Littlefield is Director, Academic Center for Excellence in Teaching; and Dr. Jones is Professor, Department of Pathology—all at the University of Texas Health Science Center at San Antonio. Direct correspondence and requests for reprints to Dr. Anne Cale Jones, Department of Pathology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229-3900; 210-567-4122 phone; 210-567-2303 fax; jonesac@uthscsa.edu.

Key words: aptitude-treatment interaction, validity, formula scoring, correction for guessing, educational methodology, educational measurement, student performance, evaluation, multiple-choice questions, short-answer questions, dental education

Submitted for publication 3/24/08; accepted 7/22/08

The expectation for students to improve their scores by guessing on multiple-choice format examinations is well known. We previously reported the results of retrospectively applying a standard correction for random (no knowledge) guessing[1-3] to the scores on multiple-choice examinations in an Oral and Maxillofacial Pathology course for dental students.[4] We found increased agreement of corrected scores from multiple-choice examinations with scores on short-answer examinations, that is, increased validity. We take as self-evident that the short-answer format examination greatly reduces the potential for guessing the correct answer. In this article, we report the results of a prospective implementation of the correction for guessing in which students were told in advance that a correction for guessing would be applied to their multiple-choice examination scores.

We wanted to accomplish three objectives by prospective implementation of a correction for guessing. The first would be improved validity for the multiple-choice examinations reflected by better

agreement of multiple-choice scores with the short-answer scores; this improved validity for the multiple-choice examinations would provide a better assessment of a student's knowledge. The second would be that a random component of the scores (luck) would be reduced, resulting in increased reliability of the multiple-choice examinations. The third objective would be a change in students' behavior toward self-recognition and acknowledgment of what they do not know. This ability to recognize and the willingness to admit what they do not know are essential attributes for future health care professionals.

Students, as well as some faculty colleagues, expressed concern that using correction for guessing would result in lower grades and thereby jeopardize a student's academic record; we shared that concern. However, the results of this study showed that students' grades were not adversely affected, but to the contrary, student performance actually improved after implementation of correction for guessing.

Methods

This study was designed to evaluate the effects of prospectively applying a correction for guessing to test scores of multiple-choice examinations. Student scores from an Oral and Maxillofacial Pathology course were compared between two academic class years. The courses in the two years were the same except that, in the second class, students were informed that correction for guessing would be employed in calculating the multiple-choice examination scores. Student course scores were derived from both short-answer and multiple-choice examinations. Scores in the prerequisite General Pathology course were used to show that the two classes were academically comparable. We statistically compared student performance in the Oral and Maxillofacial Pathology course between the two classes overall and also within grade classification in the General Pathology course to assess the effect of correction for guessing on students of different ability. Statistical analyses were also conducted to evaluate the effect of correction for guessing on the validity of multiple-choice examinations and on the reliability of examinations.

The correction for guessing that we investigated was a modification to the common grading method for multiple-choice examinations (number-correct or number-right scoring), in which 0 points are assigned for an incorrect answer and full credit is given for a correct answer.[4,5] In the multiple-choice examinations we investigated, each multiple-choice question had five possible answers. The standard correction for guessing consisted of awarding -1/4 for an incorrect answer, 0 for a question not answered, and +1 for a correct answer. The probability of guessing, assuming a random selection, the single correct answer was 1/5 (0.20), and the probability of guessing an incorrect answer was 4/5 (0.80); thus, a student was expected to have guessed an incorrect answer four times more often than he or she guessed a correct answer. Therefore, using the standard correction for guessing, the expected value of the number of points gained due to random guessing was (0.20)(1)+(0.80)(-1/4)=0. In general, for K possible answers per question, -1/(K-1) is awarded for an incorrect answer, 0 for a question not answered, and +1 for a correct answer.[5] This correction for guessing is generally referred to as formula scoring[5] or as the standard correction for guessing. Formula scoring is a special case[6] of choice weighting.

The correction for guessing was implemented prospectively in the Oral and Maxillofacial Pathology course at the University of Texas Health Science Center at San Antonio in the academic year 2005–06. The rationale and method of correction for guessing were explained in the course syllabus and by faculty who met with students during the introductory lecture to explain the procedure and answer questions. The correction for guessing in the 2004–05 Oral and Maxillofacial Pathology course was investigated retrospectively.

The results in this report are based on observations made in the Oral and Maxillofacial Pathology course in two different academic years. The only planned difference was that correction for guessing was used in the scoring of multiple-choice examinations. Any other changes between the two years, such as changes in course content or emphasis and changing or repeating of examination questions, were changes that typically occur from year to year.

The Oral and Maxillofacial Pathology course at the University of Texas Health Science Center at San Antonio, given in the spring semester to all second-year dental students, was fifty-eight hours in length and consisted of fifty hours of lecture and four two-hour examinations. Each of the four examinations was divided into two one-hour examinations.

The first hour of each examination was based on the presentation of twenty-five clinical cases. Each case consisted of a brief written clinical history and projected clinical, microscopic, and/or radiographic findings. Each student was given a written

examination consisting of the brief clinical histories corresponding to the twenty-five clinical cases that would be projected. The instructor projected the first clinical case while the students read the corresponding clinical history. The students were given several minutes to look at the projected clinical case and formulate a written response. After a period of time, as determined by the instructor, the instructor asked the class if any student needed more time before the next clinical case was projected. If even one student raised a hand, more time was given before the next clinical case was projected. When all twenty-five clinical cases were projected, the students were given any remaining time in the one-hour block to review their written responses to all of the twenty-five clinical cases and make any changes or corrections. No clinical cases were projected more than once. For each disease taught in the course, students were expected to learn the salient clinical characteristics, etiology/pathogenesis, radiographic features (if appropriate), histopathologic findings, and pertinent treatment options/prognosis. For each of the twenty-five cases in the four examinations, two short-answer questions were asked for a total of fifty questions. Students were advised to respond succinctly to the short-answer questions and not to use verbose essay-type responses. Responses to short-answer questions typically consisted of one or more sentences or several key words. The short-answer examinations were collected and subsequently graded by the course director (ACJ). The short-answer questions were graded by identifying key words delineated at the time of construction of the examination. Points were not deducted for spelling errors as long as responses were phonetically correct. If a student gave several answers, only the first answer was evaluated; no partial credit was awarded.

The second hour of each examination consisted of fifty multiple-choice questions, each with one correct answer and four plausible distractors. The multiple-choice questions were a mixture of clinical vignettes and didactic questions. Students were asked to choose the single correct answer for each question. At the end of the second hour, the multiple-choice examination answer sheets were collected and graded electronically.

Since the multiple-choice and short-answer examinations each consisted of fifty questions, they were equally weighted in the calculation of each student's final grade. Each of the four two-hour examinations comprised 25 percent of the final grade. No comprehensive final examination was given. Students received final course grades based on averages calculated from the scores on the four one-hour short-answer examinations and the four one-hour multiple-choice examinations. These averages were used to assign course letter grades as A (90–100 percent), B (80–89 percent), C (70–79 percent), or F (0–69 percent). These arbitrary but commonly used grade cutpoints are used throughout this report to classify students.

Each of the four examinations covered between eleven and thirteen hours of lecture material. When the individual short-answer and multiple-choice examinations were constructed, the questions were equally weighted to the topics that were presented prior to each of the four examinations. This was to ensure that a given topic was not stressed more often than another topic. The students were advised to add up the number of topics discussed in a given section and divide that number by 50 to arrive at an approximate number of questions per topic on both the multiple-choice and short-answer examinations.

The General Pathology course at the University of Texas Health Science Center at San Antonio, given in the fall semester to all second-year dental students, immediately precedes the Oral and Maxillofacial Pathology course. The course was seventy-seven hours in length and consisted of sixty-one hours of lecture, four two-hour review sessions, and four two-hour examinations. The review sessions were structured in a question and answer format. Each faculty member who had previously presented didactic information for the upcoming examination presented a brief verbal review of his or her topic. Students were then allowed to ask questions, and topics were discussed in further detail by the faculty member. This procedure was repeated until there were no further questions. Each of the examinations consisted of seventy-five multiple-choice questions with one correct answer and four distractors; test construction strategies were similar to those described for the Oral and Maxillofacial Pathology course. The multiple-choice questions covered information presented in the lectures and reading assignments. Each two-hour examination comprised 25 percent of the final course grade. No comprehensive final examination was given. Students received a final course grade based on the averages calculated from the four two-hour examinations. These averages were used to assign course grades using the same categories as in the Oral and Maxillofacial Pathology course. Each of the four examinations covered between thirteen and nineteen hours of lecture material. When the multiple-choice ques-

tions were constructed, the questions were equally weighted to the topics presented prior to each of the four examinations. This was to ensure that a given topic was not stressed more often than another. The students were advised to add up the number of topics discussed in a given section and divide that number by 75 in order to arrive at an approximate number of questions per topic.

Ninety students initially enrolled in the Oral and Maxillofacial Pathology course during the 2004–05 academic year; two students who were failing the course after the completion of three examinations withdrew from school before the fourth examination. Eighty-eight students were enrolled in the Oral and Maxillofacial Pathology course in 2005–06. Three students who were repeating the course, one student who did not take the third examination, and one student who took General Pathology out of sequence were excluded from the 2005–06 class. Thus, the analyses presented in this report were based on eighty-eight students (2004–05) and eighty-three students (2005–06) who completed the General Pathology course prior to enrolling in the Oral and Maxillofacial Pathology course and who subsequently had scores for all four examinations in the Oral and Maxillofacial Pathology course. This study was approved by the Institutional Review Board of the University of Texas Health Science Center at San Antonio.

Means were compared between classes using the t-test for independent samples, and standard deviations (variances) were compared using the F-test.[7] If variances were unequal, Satterthwaite's modification to the t-test was used.[8] Cumulative relative frequency distributions were used to graphically show distributions of numerical scores. Frequencies in letter grade groups (A, B, C, or F) were analyzed using the chi-square test for independence.[7] Where low expected frequencies were encountered, an exact procedure was used to obtain the P-value. We classified students based on their letter grade (A, B, C, F) in the General Pathology course (a measure of student aptitude) and then analyzed their numerical scores in the Oral and Maxillofacial Pathology course by these classifications and class year (retrospective or prospective correction for guessing, that is, the treatment) in a two-way analysis of variance[7] to investigate aptitude-treatment interactions.[9]

Aggregate agreement[10] of scores from multiple-choice examinations and scores from short-answer examinations was assessed using principal component lines[11] estimated from the variance-covariance matrix. The first principal component is the line through the means $(\overline{X}, \overline{Y})$, which minimizes the sum of the squared distances of the data points to the line.[12] We used principal components analysis because both the X and Y variables were random variables. In linear regression analysis, only the Y variable is considered to be a random variable, and the estimator of the line is biased if X also is a random variable.[13] Thus, the first principal component lines more accurately estimate the relation between these X and Y variables.

A bootstrap procedure,[14] with 1,000 samples, was used to estimate confidence intervals for the slope and intercept of the first principal component lines, to test that the slope was 1.00 and the intercept 0.00, and to test equality of the principal component lines for the 2004–05 and 2005–06 classes. The bootstrap is a nonparametric procedure and thus does not depend on any particular probability distribution. The statistic of interest is calculated in bootstrap samples, of the same size as the original, that are generated by sampling with replacement from the original data. Thus, the bootstrap is a resampling procedure. If the resampling is repeated a large number of times, the empirical distribution of the statistic generated from many bootstrap samples approximates the actual distribution. The empirical distribution may be used to construct confidence intervals (95 percent confidence limits are the 2.5 and 97.5 percentiles of the empirical distribution) or perform hypothesis tests.

The Cronbach's alpha statistic was used as a measure of reliability. Carmines and Zeller[15] describe the Cronbach's alpha statistic as an estimate of the expected correlation between one test and a hypothetical alternative form containing the same number of items.

Results

To determine whether the 2004–05 and 2005–06 classes were academically comparable, we compared their respective performances in the General Pathology course taken in the semester immediately preceding the Oral and Maxillofacial Pathology course. The distributions of individual student course scores are shown in Figure 1. This figure shows the cumulative relative frequencies, that is, the fraction of students at a particular grade average and below. Little difference in distributions of grade averages between the 2004–05 and 2005–06 classes in the General Pathology course is indicated. This lack of

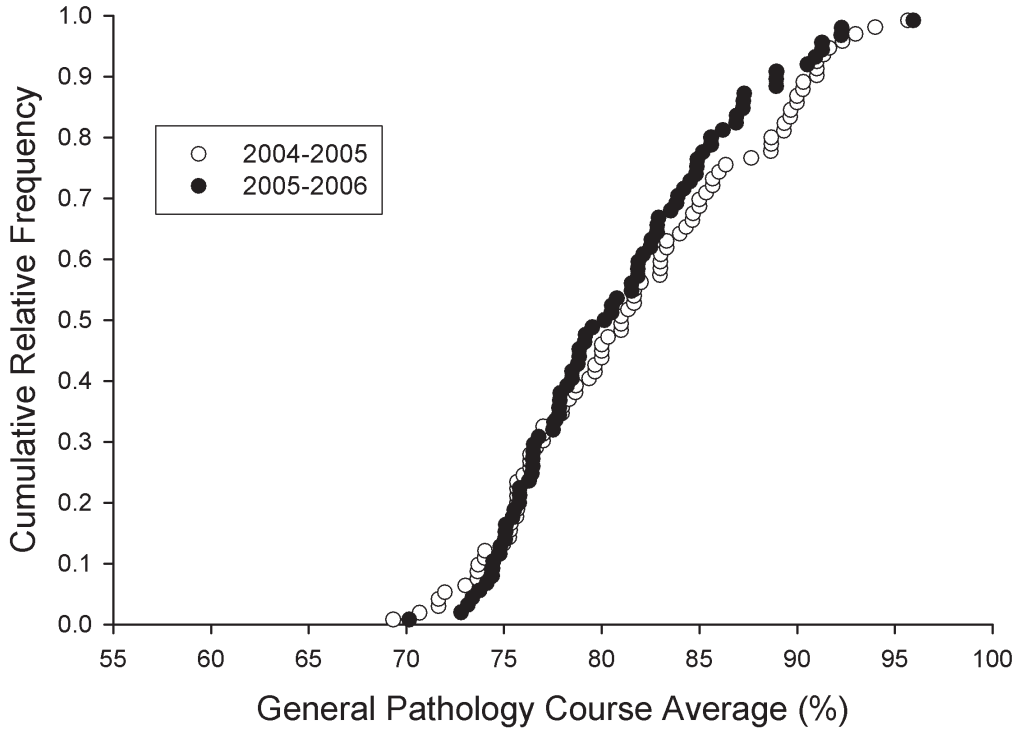

Figure 1. General Pathology course scores

Note: Shown are cumulative relative frequency distributions of student averages from four multiple-choice examinations in the General Pathology course in 2004–05 and 2005–06.

difference was further reflected quantitatively; the class means were 81.6 ±6.4 (SD) in 2004–05 and 80.9 ±5.6 in 2005–06 (P=0.4369). Further analysis also showed there was no significant difference in letter grade distributions (A, B, C, or F) between the 2004–05 and 2005–06 classes (Table 1); that is, the fractions of students receiving A, B, C, or F in the two classes were similar.

The distributions of individual student scores on the short-answer and multiple-choice examinations for the Oral and Maxillofacial Pathology course are shown in Figure 2. This figure shows that, after correction for guessing was implemented prospectively, the lower part of the distributions of both short-answer and multiple-choice scores was clearly higher (curve shifted to the right) in the 2005–06 class compared to the 2004–05 class. The class means of the short-answer examination scores were 82.8 ±8.6 (SD) in 2004–05 and 84.8 ±6.4 in 2005–06 (P=0.0895); and the class means of the multiple-choice examination scores were 81.6 ±7.7 in 2004–05 (retrospectively corrected for guessing) and 83.5 ±6.1 in 2005–06 (P=0.0722). The resultant overall course means were 82.2 ±7.7 in 2004–05 and 84.1 ±5.6 in 2005–06 (P=0.0607). The standard deviations of short-answer examination scores, multiple-choice examination scores, and overall course

Table 1. Grade distribution (%) in the General Pathology course by year

Grade	2004–05 Class	2005–06 Class	Significance
A	14.8%	8.4%	P=0.3941
B	42.1%	42.2%	
C	42.1%	49.4%	
F	1.1%	0.0%	

Note: Percentages for 2004–05 do not total 100% because of rounding.

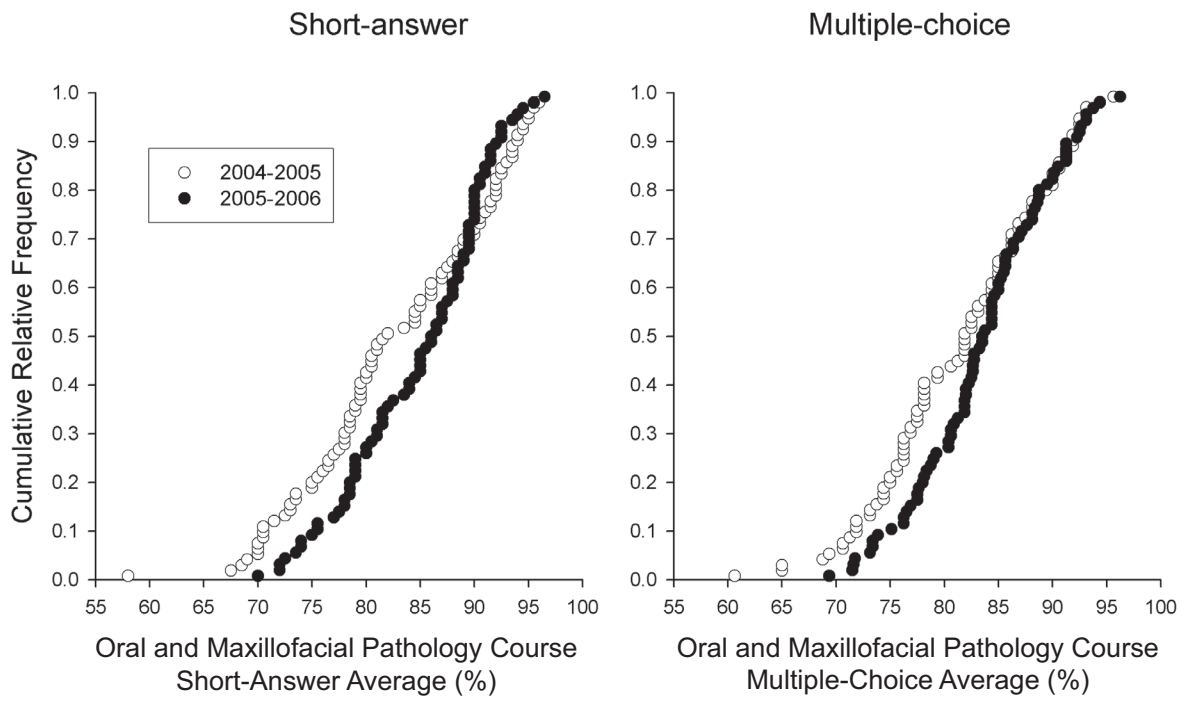

Figure 2. Oral and Maxillofacial Pathology course averages

Note: Shown are cumulative relative frequency distributions of student averages from four short-answer examinations (left panel) and four multiple-choice examinations (right panel) in the Oral and Maxillofacial Pathology course in 2004–05 and 2005–06. Multiple-choice examination scores in 2004–05 were retrospectively corrected for guessing.

scores were significantly smaller in 2005–06 than in 2004–05 (P≤0.0387). The shift to higher grades by the lower-performing students noted in Figure 2 was responsible for both the higher means and the smaller standard deviations in the 2005–06 class relative to the 2004–05 class.

The upward shift in the lower part of the grade distribution in 2005–06 was further examined by comparing the fraction of students in each of the grade categories A, B, C, and F between the two classes (Table 2). There was a significant difference in overall grade (average of multiple-choice and short-answer scores) distribution (P=0.0100), in short-answer grade distribution (P=0.0190), and in multiple-choice grade distribution (P=0.0559), between the 2004–05 and 2005–06 classes. Of note, for overall grade, short-answer grade, and multiple-choice grade there were significantly (P≤0.0305) lower fractions of Cs and Fs and increased fractions of As and Bs, in 2005–06 compared to 2004–05. Clearly, prospective implementation did not adversely affect the students' grades, but, in fact, resulted in improved student performance in 2005–06 relative to 2004–05.

The relationships between an individual student's General Pathology course score and his or her subsequent scores on the short-answer and multiple-choice examinations in the Oral and Maxillofacial Pathology course are shown in Figure 3. The "C" students in the General Pathology course are shown in the shaded areas. Clearly, the subsequent performances in the Oral and Maxillofacial Pathology course by the 2005–06 "C" General Pathology students were higher than the performances of the 2004–05 "C" General Pathology students; this is indicated by the observation that the number of filled circles (representing grades for 2005–06 students) in the shaded area that shifted to A or B was much greater than the number of open circles (representing grades for 2004–05 students) that shifted higher to A or B. To verify this visual impression, we analyzed the numerical scores in the Oral and Maxillofacial Pathology course between the two classes with stu-

Table 2. Grade distribution (%) in the Oral and Maxillofacial Pathology course by examination format and year

Examination Format	Grade	2004–05 Class	2005–06 Class	Significance
Overall*	A	21.6%	19.3%	P=0.0100
	B	37.5%	57.8%	
	C	35.2%	22.9%	
	F	5.7%	0.0%	
Short-answer	A	29.6%	26.5%	P=0.0190
	B	29.6%	48.2%	
	C	36.4%	25.3%	
	F	4.6%	0.0%	
Multiple-choice*	A	19.3%	18.1%	P=0.0559
	B	37.5%	55.4%	
	C	37.5%	25.3%	
	F	5.7%	1.2%	

*Retrospectively corrected in 2004–05.
Note: Percentages may not total 100% because of rounding.

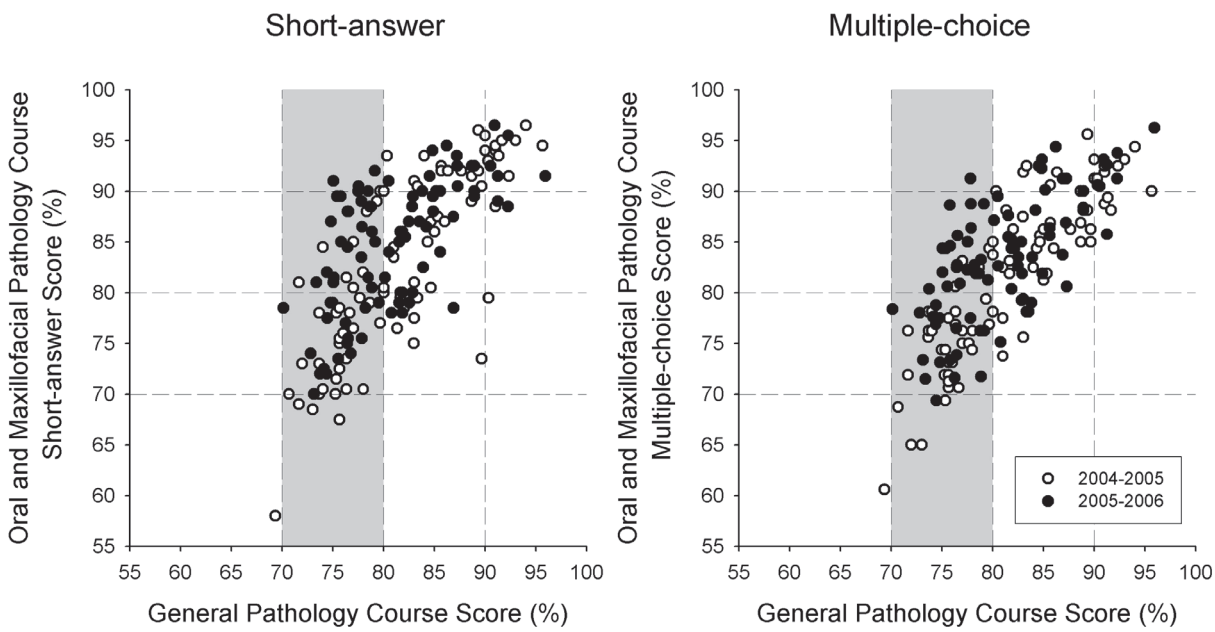

Figure 3. Oral and Maxillofacial Pathology course scores vs. General Pathology course scores

Note: Shown are scatterplot of average score on four short-answer examinations in the Oral and Maxillofacial Pathology course by average score on four multiple-choice examinations in the General Pathology course during academic years 2004–05 and 2005–06 (left panel) and scatterplot of average score on four multiple-choice examinations in the Oral and Maxillofacial Pathology course by average score on four multiple-choice examinations in the General Pathology course during academic years 2004–05 (retrospectively corrected) and 2005–06 (right panel). Grid lines indicate commonly used grade cutpoints of 70, 80, and 90 percent. Shaded area identifies students with a grade of C (70–79 percent) in the General Pathology course.

dents classified by their letter grade in the General Pathology course. These analyses showed significant aptitude-treatment interactions[9] between the letter grade in the General Pathology course and class year (short-answer, P=0.0447; multiple-choice, P=0.0246; and course average, P=0.0113). Table 3 shows that those students with a C in the General Pathology course had higher average scores (short-answer, multiple-choice, and overall) in the Oral and Maxillofacial Pathology course in the 2005–06 class compared to the 2004–05 class. This significant difference is reflective of an aptitude-treatment interaction because the difference between classes was observed only in the C students and not in the A or B students—that is, the improvement depends on the student's ability or aptitude as measured by the grade in the General Pathology course. Furthermore, in 2004–05, 37.8 percent (14/37) of students with a C in the General Pathology course subsequently received a B (no As) in the Oral and Maxillofacial Pathology course, whereas in 2005–06, 58.5 percent (24/41) of students with a C in the General Pathology course received an A or B in the Oral and Maxillofacial Pathology course (P=0.0678). Of the students receiving a B in the General Pathology course, the fractions that improved to an A in the Oral and Maxillofacial Pathology course were similar in the two classes (33.3 percent [12/36] in 2004–05; 27.3 percent [9/33] in 2005–06; P=0.5847).

We were concerned that students in the 2005–06 Oral and Maxillofacial Pathology course might have responded to the imposition of the correction for guessing by seeking information about examinations from prior classes. In this regard, examining the grades that the students achieved on the same questions as well as different questions in the two years showed

Table 3. Mean scores in the Oral and Maxillofacial Pathology course by grade category in the General Pathology course, year, and examination format

	General Pathology		Oral and Maxillofacial Pathology					
	Number of students		Short-Answer		Multiple-Choice		Overall	
	2004–05	2005–06	2004–05	2005–06	2004–05†	2005–06	2004–05†	2005–06
			Mean ±SE	Mean ±SE	Mean ±SE	Mean ±SE	Mean ±SE	Mean ±SE
A	13	7	92.7 ±1.2	92.1 ±1.1	91.3 ±0.5	91.9 ±1.2	92.0 ±0.7	92.0 ±1.0
B	37	35	85.9 ±1.0	86.7 ±0.8	84.9 ±0.8	85.7 ±0.8	85.4 ±0.8	86.1 ±0.7
C	37	41	77.0 ±1.0	81.9 ±1.0*	75.4 ±0.8	80.2 ±0.8*	76.2 ±0.8	81.0 ±0.8*
F	1	0	58		60.6		59.3	

*2005–06 significantly different (P≤0.05) from 2004–05.
†Retrospectively corrected for guessing.

Table 4. Means of scores on short-answer and multiple-choice examinations for the same questions and different questions used in the Oral and Maxillofacial Pathology course, by year

			Same Questions	Different Questions
Examination Format	Class	Number of Students	Mean ±SE	Mean ±SE
Short-answer*			n=113†	n=87†
	2004–05	88	82.7 ±0.98	82.6 ±0.97
	2005–06	83	85.4 ±0.80	84.4 ±0.74
			P=0.0389	P=0.1472
Multiple-choice			n=115†	n=85†
	2004–05	88	83.8 ±0.93‡	77.8 ±0.92‡
	2005–06	83	84.8 ±0.78	80.5 ±0.91
			P=0.4108	P=0.0359

*Eighty-six of the 100 clinical cases used in the short-answer examinations were the same for the 2004–05 and 2005–06 classes.
†Number of questions.
‡Retrospectively corrected for guessing.

that the improved performance by the 2005–06 class in the Oral and Maxillofacial Pathology course was not due to information about the examination questions obtained from prior classes (Table 4).

One objective was to assess the effect of the correction for guessing on the validity of the multiple-choice examinations. The agreement of the scores of the individual 2005–06 students on the multiple-choice examinations with the scores on the short-answer examinations in the Oral and Maxillofacial Pathology course is shown graphically in Figure 4; and the equation for the first principal component line, describing aggregate agreement, is given in Table 5. The line for the trend in Figure 4 is close to the line of equality, indicating good validity of the corrected multiple-choice examination scores. As shown in Table 5, the slope was not significantly different from 1 and the intercept was not significantly different from 0. Also given in Table 5 is the equation for the first principal component line previously reported[4] from the retrospective application of the correction for guessing. The slope of the line obtained after prospective implementation of correction for guessing was closer to 1 and the intercept closer to 0 than was the line obtained after retrospective application of correction for guessing; however, these differences were not statistically significant.

For the Oral and Maxillofacial Pathology course, the Cronbach's alpha statistics, measuring examination reliabilities, for the multiple-choice examinations was 0.89 in 2004–05 and 0.87 in 2005–06; for short-answer examinations, the Cronbach's alpha statistics were 0.89 in 2004–05 and 0.87 in 2005–06. Thus, reliabilities of the examinations in the 2004–05

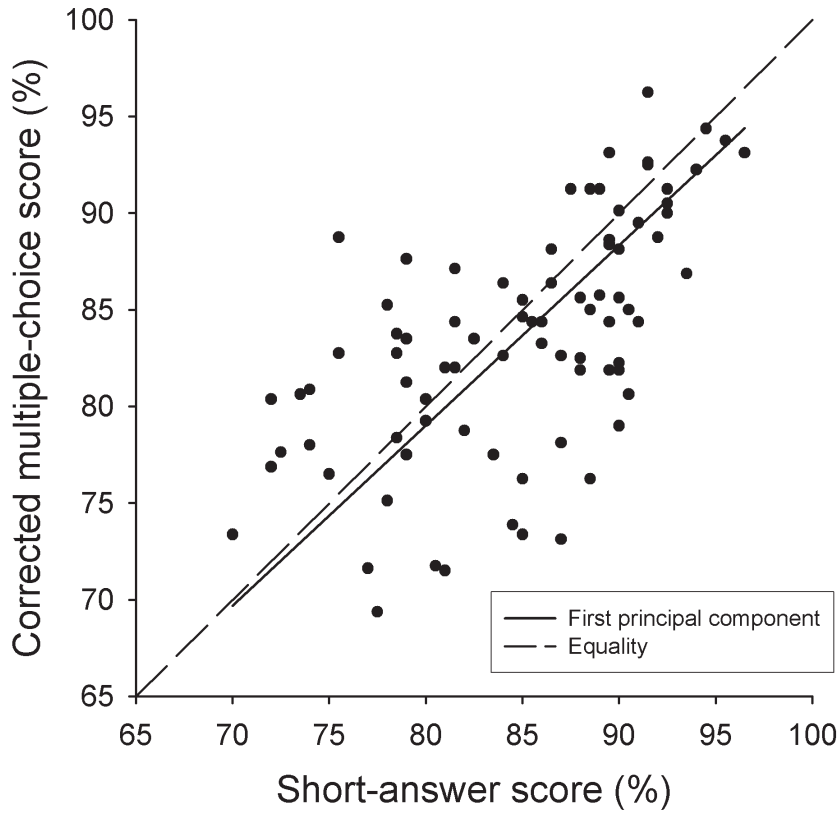

Figure 4. Oral and Maxillofacial Pathology course, 2005–06

Note: Shown is scatterplot of average scores from four multiple-choice examinations and scores from four short-answer examinations in the Oral and Maxillofacial Pathology course in academic year 2005–06. The solid line represents the first principal component and the dashed line represents equality. The slope of the first principal component was not significantly different from 1.00 and the intercept not significantly different from 0.00 (see Table 5).

Table 5. Slope and intercept for equation of the first principal components between corrected multiple-choice scores and short-answer scores in the Oral and Maxillofacial Pathology course, by class

Class	First Principal Component Slope (95% CI)	Intercept (95% CI)
2004–05*	0.87 (0.76 to 0.98)	9.51 (0.06 to 18.50)
2005–06	0.93 (0.72 to 1.23)	4.42 (-21.36 to 22.32)
Significance	P=0.568	P=0.618

*From Prihoda TJ, Pinckard RN, McMahan CA, Jones AC. Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. J Dent Educ 2006;70(4):378–86.

and 2005–06 classes were adequate (≥0.80) as defined by Carmines and Zeller.[15]

An objective of using the correction for guessing was to encourage students to recognize and admit what they did not know. Eighty-nine percent (88.9 percent) of students in the 2005–06 Oral and Maxillofacial Pathology course with numerical course scores of 70–79 percent left at least one multiple-choice question unanswered, 56.3 percent of students with scores of 80–89 percent left at least one question unanswered, and 26.7 percent of students with scores of 90–100 percent left at least one question unanswered. Students with course scores of 70–79 percent left a mean of 3.50 multiple-choice questions unanswered (from a total of 200 multiple-choice questions), while students with course scores of 80–89 percent left 2.02 multiple-choice questions unanswered and students with 90–100 percent left 0.60 questions unanswered. The average fractions of questions that the student did not know that were left unanswered (percentage of question left unanswered as a fraction of number left unanswered plus number answered incorrectly) were 9.0 percent, 6.9 percent, and 4.0 percent for students with numerical course scores of 70–79 percent, 80–89 percent, and 90–100 percent, respectively.

Discussion

We were surprised and pleased to learn that lower-performing "C" students in our General Pathology course significantly improved their grades in the Oral and Maxillofacial Pathology course after we prospectively implemented the correction for guessing. Our interpretation of this aptitude-treatment interaction[9] is that the improved performance, in both the short-answer and multiple-choice examinations, was because the 2005–06 students were concerned about the consequences of the correction for guessing, and as a result, they studied more diligently. In doing so, they attained a better overall understanding of the subject matter. Thus, the improved performance represented a positive (behavioral) attitude of the 2005–06 students in response to our raising of the bar.[16]

We arrived at the foregoing interpretation of the reason for the improved student performance after implementation of correction for guessing after excluding other potential explanations. The comparability of the examination scores and final grades between the two classes in our General Pathology course supports the premise that the two classes were not academically different and that such a difference did not account for the improved performance in the Oral and Maxillofacial Pathology course by the 2005–06 students. The improved performance on both the same questions and different questions suggests that improved performance in 2005–06 apparently was not because examination questions were passed along from previous classes.

Our objective that students would not only recognize but acknowledge what they do not know was, at best, only modestly achieved as indicated by the relatively small number and fraction of questions left unanswered. There are several potential explanations for these findings. The standard correction for guessing adjusted only for truly random guessing among five possible answers. Thus, it potentially would benefit students to guess if they could eliminate one, two, or three of the distractors, which has been a concern with multiple-choice questions.[17] Prihoda et al.[4] previously presented the expected gain per question if a student had partial knowledge and could eliminate one or more of the distractors. One student in the 2005–06 class publicly advised his classmates to continue to guess; his argument was based on expected gain and he did not consider the probability of students guessing themselves into a lower grade.

There was a slight improvement in agreement between multiple-choice and short-answer scores but little or no change in reliability. The principal component line (that is, the single dimension that best summarizes the data from both multiple-choice and short-answer examinations) was closer, although not significantly closer, to the line of equality after prospective implementation of the correction for guessing than was the line reported by Prihoda et

al.[4] (Table 5) after retrospective correction. These results continue to support increased validity[18] due to applying the standard correction for guessing to multiple-choice examination scores. Our use of validity refers to performance without "cuing" in the short-answer examination. While we cannot claim that the short-answer examination better evaluates student knowledge based on these data only, we believe this question format, which reduces the influence of guessing, will be a better indicator of what students know or do not know about a given subject. Furthermore, the short-answer question format as used in the Oral and Maxillofacial Pathology course not only tests knowledge of memorized facts,[19] but also requires critical thinking to integrate clinical characteristics, radiographic features, and histopathologic findings to arrive at an appropriate diagnosis of various disease processes. We are convinced that a short-answer examination provides a better measure of a student's ability to perform in clinical situations than does a multiple-choice examination.

The correction for guessing could be implemented in any course that uses multiple-choice examinations; implementation would be most effective if used in all applicable courses within an institution. In order for the numerical value subtracted for incorrect responses in applying the correction for guessing to be accurate, it is imperative that functional distractors be used, that is, students must not be able to easily eliminate incorrect answers. The short-answer format examination can be used in any didactic course although grading such examinations requires extensive effort by the faculty.

Conclusions

Prospective implementation of correction for guessing in multiple-choice examinations in an Oral and Maxillofacial Pathology course resulted in significantly better student performance as indicated by improvements in numerical scores and by letter grade distribution on both short-answer and multiple-choice examinations. Moreover, prospective implementation of correction for guessing resulted in improved validity of multiple-choice examinations. This study supports the premise that these health professions students, qualified through a competitive selection process, respond positively to an increase in expectation for student performance and that health professions faculty should not be reluctant to raise the bar for them.

Acknowledgment

The authors gratefully acknowledge Ms. Belen Ballesteros for her excellent management of the database of student test scores that were used in this study.

REFERENCES

1. Diamond J, Evans W. The correction for guessing. Rev Educ Res 1973;43:181–91.
2. Rogers HJ. Guessing in multiple choice tests. In: Masters GN, Keeves JP, eds. Advances in measurement in educational research and assessment. New York: Pergamon, 1999.
3. Muijtjens AMM, van Mameren H, Hoogenboom RJI, Evers JLH, van der Vleuten CPM. The effect of a "don't know" option on test scores: number-right and formula scoring compared. Med Educ 1999;33:267–75.
4. Prihoda TJ, Pinckard RN, McMahan CA, Jones AC. Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. J Dent Educ 2006;70(4):378–86.
5. Lord FM. Formula scoring and number-right scoring. J Educ Measurement 1975;12:7–12.
6. Davis FB, Fifer G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educ Psychol Measurement 1959;19:159–70.
7. Snedecor GW, Cochran WG. Statistical methods. Ames: Iowa State University Press, 1967.
8. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics 1946;2:110–4.
9. Snow R. Aptitude-treatment interaction as a framework for research on individual differences in learning. In: Ackerman P, Sternberg RJ, Glaser R, eds. Learning and individual differences. New York: W.H. Freeman, 1989.
10. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;327:307–10.
11. Morrison DF. Multivariate statistical methods. Belmont, CA: Brooks/Cole Thomson Learning, 2005.
12. Seber GAF. Multivariate observations. New York: Wiley, 1984.
13. Draper NR, Smith H. Applied regression analysis. New York: John Wiley & Sons, 1998.
14. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall, 1993.
15. Carmines EG, Zeller RA. Reliability and validity assessment. London: Sage Publications, 1979.
16. Natriello G, Dornbusch SM. Teacher evaluative standards and student effort. New York: Longman, 1984.
17. Haladyna TM. Developing and validating multiple-choice test items. 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2004.
18. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ 2003;37:830–7.
19. Gronlund NE, Linn RL. Measurement and evaluation in teaching. New York: Macmillan, 1990.