Improving Multiple-Choice Questions to Better Assess Dental Student Knowledge: Distractor Utilization in Oral and Maxillofacial Pathology Course Examinations

C. Alex McMahan, Ph.D.; R. Neal Pinckard, Ph.D.; Thomas J. Prihoda, Ph.D.; William D. Hendricson, M.A., M.S.; Anne Cale Jones, D.D.S.

Abstract: How many incorrect response options (known as distractors) to use in multiple-choice questions has been the source of considerable debate in the assessment literature, especially relative to influence on the likelihood of students’ guessing the correct answer. This study compared distractor use by second-year dental students in three successive oral and maxillofacial pathology classes that had three different examination question formats and scoring resulting in different levels of academic performance. One class was given all multiple-choice questions; the two other were given half multiple-choice questions, with and without formula scoring, and half un-cued short-answer questions. Use by at least 1 percent of the students was found to better identify functioning distractors than higher cutoffs. The average number of functioning distractors differed among the three classes and did not always correspond to differences in class scores. Increased numbers of functioning distractors were associated with higher question discrimination and greater question difficulty. Fewer functioning distractors fostered more effective student guessing and overestimation of academic achievement. Appropriate identification of functioning distractors is essential for improving examination quality and better estimating actual student knowledge through retrospective use of formula scoring, where the amount subtracted for incorrect answers is based on the harmonic mean number of functioning distractors.

Dr. McMahan and Dr. Pinckard contributed equally to this study. Dr. McMahan is Professor, Department of Pathology, University of Texas Health Science Center at San Antonio; Dr. Pinckard is Distinguished Teaching Professor, Department of Pathology, University of Texas Health Science Center at San Antonio; Dr. Prihoda is Professor, Department of Pathology, University of Texas Health Science Center at San Antonio; Mr. Hendricson is Assistant Dean, Educational and Faculty Development, Dental School, University of Texas Health Science Center at San Antonio; and Dr. Jones is Distinguished Teaching Professor, Department of Pathology, University of Texas Health Science Center at San Antonio. Direct correspondence and reprint requests to Dr. C. Alex McMahan, Department of Pathology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229-3900; 210-567-4026; mcmahan@uthscsa.edu.

Keywords: assessment, multiple-choice questions, functioning distractors, formula scoring, correction for guessing, educational assessment, dental education, oral and maxillofacial pathology

Submitted for publication 11/27/12; accepted 1/20/13

In our study published in 2012, we documented that two of many aspects of the learning environment—examination question format and scoring procedure—had significant but separate effects on the grades of second-year dental students in an oral and maxillofacial pathology course.¹ Relative to all multiple-choice questions, the use of short-answer questions in half of the examinations significantly improved course scores, including scores on the multiple-choice portions.¹,² Furthermore, when the multiple-choice portions of the examinations were scored by prospectively implementing formula scoring, there were significant additional increases in academic performance by lower performing students; of note, this enhanced academic performance was observed not only on the multiple-choice questions but also on the short-answer portions.¹,³ In contrast, when the examinations consisted of all short-answer questions, there was no significant difference in academic performance compared to the class in which examinations were comprised of half short-answer questions and half multiple-choice questions scored without formula scoring.¹,²

It is widely acknowledged that multiple-choice examinations overestimate student academic achievement. This is because multiple-choice questions are cued and students are rewarded for guessing the correct answer; moreover, the likelihood of attaining this reward for guessing is higher when students can eliminate one or more incorrect options. Indeed, we previously reported that, by using the formula scoring algorithm, the scores of individual students on the multiple-choice portions of their examinations agreed significantly better with their scores on
the short-answer portions of examinations testing the same body of material; therefore, we concluded that correction for guessing increased the validity of scores obtained on the multiple-choice portions.\(^3\),\(^4\)

In those studies, the formula scoring algorithm we used was based on random, no knowledge guessing among five possible options. Thus, if many students did exclude one or more of the incorrect distractors, our previously published assessments of score inflation on multiple-choice examinations were underestimated.

A “distractor” is an incorrect response option that students must take into consideration when determining their answer on a multiple-choice question (MCQ). “Functioning distractors” are those actually involved in the student’s response selection process. Operationally, Haladyna and Downing\(^5\) defined a functioning distractor as one whose use increased as student scores decreased; that discriminated between lower and higher performing students; and that was selected by at least 5 percent of students. However, recently Rogausch et al.\(^6\) suggested using 1 percent of students as a cutoff value; their view was that distractors could not be expected to attract more than 1 percent of well-prepared examinees.

The question of how many distractors to use in MCQs has been the source of considerable debate, as has the influence the number of functioning distractors has on students’ ability to guess the correct response.\(^5\) Developing appropriate modalities to define and enumerate the numbers of functioning distractors is an essential element of examination and course improvement programs as well as accurate assessment of students’ knowledge; therefore, assisting with that process was the aim of our study. We began by estimating the number of functioning distractors in MCQs comprised of one correct answer and four incorrect options. We subsequently determined whether the number of functioning distractors varied among different learning and examination environments, as we previously documented for student course score.\(^1\),\(^3\)

We anticipated that the numbers of distractors we identified as functioning would be associated with question difficulty and question discrimination. Following Haladyna and Downing,\(^5\) we required that a functioning distractor have a frequency of use (number or fraction of students) equal to or greater than a defined cutoff value. We investigated the effects of cutoff values of 1 percent\(^6\) to 5 percent\(^7\) of students on the estimated number of functioning distractors used by our students in an oral and maxillofacial pathology course. We then calculated the appropriate number of distractors to use in the formula scoring algorithm to adjust for guessing and the inflation of scores on multiple-choice exams. Estimation of the numbers of functioning distractors would enable faculty members to determine how well their exams are meeting the targeted number of functioning distractors for each question and identify needed question modifications. Furthermore, we believe that faculty members should use retrospective application of formula scoring, based on the estimated numbers of functioning distractors, to more accurately assess the actual level of students’ knowledge\(^4\) and how the particular learning environment affects their academic achievement.

---

**Methods**

**Classes in Study**

This study was approved by the Institutional Review Board of the University of Texas Health Science Center at San Antonio. The numbers of functioning distractors in multiple-choice questions were compared among three classes of second-year dental students in an oral and maxillofacial pathology course. Course content, including clinical cases as well as the lecturers, was unchanged in the three years; however, there were differences in examination question format (short-answer [SA] or multiple-choice [MC]) and the scoring algorithm used to calculate multiple-choice exam scores (number-correct scoring or formula scoring [MC\(^*\)]). Question formats, scoring methods, and abbreviations used to identify the classes are briefly summarized in Table 1. We previously detailed the history and reasons for changing exam question format and scoring procedure.\(^2\) The oral and maxillofacial pathology course at the University of Texas Health Science Center at San Antonio Dental School was presented in the spring semester of the second year of the curriculum. Analysis of course scores in the prerequisite general pathology course, presented in the preceding fall semester, showed that the three classes were academically comparable.\(^1\)

The oral and maxillofacial pathology course consisted of fifty hours of didactic lectures and four two-hour examinations. Students were informed of exam question format, scoring procedures, and calculation of final course scores in the written course syllabus, as well as being informed orally by the course director during the first class. Students
were informed explicitly, in the written syllabus and verbally, that they were expected to learn and understand the following characteristics for each of the pathologic processes discussed in the course: etiology, pathogenesis, age and sex predilection, most common anatomic location, distinguishing features (clinical, radiographic, microscopic), diagnostic aids and laboratory tests, treatment options, and prognosis. Both the short-answer and the multiple-choice exams were constructed to test the students’ knowledge and understanding of these specific disease characteristics.

Each of the four examinations in the course assessed students’ comprehension of material presented in the preceding eleven to thirteen hours of lecture. The questions on both short-answer and multiple-choice examinations were equally weighted to the topics presented prior to each of the four exams. This ensured that each covered a broad range of topics and that a given topic was not stressed more than another.

In the 2004-05 (SA/MC) and 2005-06 (SA/MC*) classes, each of the four examinations was divided into two back-to-back one-hour sessions. The first hour of each examination consisted of fifty questions based on twenty-five clinical cases. Each case consisted of a brief written clinical history and projected gross, microscopic, and/or radiographic findings. Students were told to respond succinctly to two short-answer questions for each case; appropriate answers to these questions typically consisted of one or more sentences or key words. All short-answer exams were graded solely by the course director by identifying key words delineated at the time the exam was constructed. The short-answer questions were collected at the conclusion of the first hour of the exam.

The second hour of each examination in the 2004-05 (SA/MC) and 2005-06 (SA/MC*) classes consisted of fifty multiple-choice questions, each having five options (one correct answer and four incorrect options). In the 2005-06 (SA/MC*) class, the students were informed in the written syllabus and verbally by the course director during the first class that a scoring procedure employing a correction for random, no knowledge guessing would be applied to the multiple-choice examinations.

In the 2006-07 (MC) class, each of the four two-hour examinations was comprised of seventy-five multiple-choice questions of the same nature as for the 2004-05 and 2005-06 classes. The MCQs given to the 2006-07 class covered the same topics as the combination of short-answer questions and MCQs used in 2004-05 and 2005-06 classes. The clinical cases included in the examinations in 2006-07 were the same as the cases used in the examinations in 2004-05 and 2005-06.

Students received a final course grade based on the average scores calculated from the four two-hour examinations. The reliabilities of the exams, measured by the Cronbach’s alpha statistic, were reported previously. All reliabilities were greater than 0.80, the desirable level advocated by Carmines and Zeller. Graded exams were not returned to students and were stored securely; however, students, one at a time, were permitted to review their exam for a maximum of twenty minutes under the strict supervision of a staff member of the Department of Pathology; taking of notes and use of references by students were not allowed.

### Table 1. Examination formats and scoring methods used in oral and maxillofacial pathology course, by class

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of Students</th>
<th>Fraction of Course Score</th>
<th>Number of Questions</th>
<th>Formula Scoring</th>
<th>Fraction of Course Score</th>
<th>Number of Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004-05 (SA/MC)</td>
<td>88</td>
<td>50%</td>
<td>200</td>
<td>Retrospective¹</td>
<td>50%</td>
<td>200</td>
</tr>
<tr>
<td>2005-06 (SA/MC*)</td>
<td>82</td>
<td>50%</td>
<td>200</td>
<td>Prospective</td>
<td>50%</td>
<td>200</td>
</tr>
<tr>
<td>2006-07 (MC)</td>
<td>88</td>
<td>100%</td>
<td>300</td>
<td>Retrospective¹</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

SA=short-answer questions; MC=multiple-choice questions; MC*=multiple-choice questions with prospective application of formula scoring

¹Formula scoring was not used in computation of official grades; number-correct scoring was used for official grades.

### Number of Functioning Distractors

We estimated the number of functioning distractors in multiple-choice questions comprised of the correct answer and four incorrect options. Haladyna and Downing recommended that the criterion for a functioning distractor was use by at least 5 percent of
students, while Rogausch et al. suggested use by at least 1 percent of students as a cutoff value.

The maximum use of each distractor in a question with four functioning distractors occurs when the incorrect responses are distributed uniformly over the distractors; that is, the use of each distractor is about one-fourth of the number of incorrect answers. Thus, for a question that 90 percent of students answer correctly, a question with four functioning distractors will have the maximum use of all distractors if the likelihood of use is 2.5 percent for each distractor. For more difficult questions with four functioning distractors where 70 to 80 percent of students answer correctly, the maximum use of each distractor corresponds to likelihoods of use of 7.5 and 5 percent, respectively, for each of the four distractors. For a question having three functioning distractors, the maximum use of each of three distractors is one-third of the number of incorrect answers, respectively, for each of the four distractors. For a question having three functioning distractors, the maximum use of each of three distractors is one-half or all of the number of the incorrect answers, respectively, and the other distractors are not used by any student.

We described above a question answered correctly by 90 percent of students with each of four functioning distractors used by 2.5 percent of students. If this same 2.5 percent chance of use occurred in a question with three, two, or one functioning distractors, the associated fraction of correct answers must be 92.5 percent, 95 percent, and 97.5 percent, respectively. Thus, a constant chance of distractor use resulted in questions of decreasing difficulty as the number of functioning distractors was decreased. This association of number of functioning distractors with question difficulty was part of Haladyna and Downing’s definition of functioning distractors.

Using a probability model, we investigated the use of the cutoff values of 1 percent to 5 percent of students in determining the number of functioning distractors. Our probability model assumed that all the multiple-choice questions were comprised of one correct answer and four incorrect options. The probability of a student giving the correct answer to the jth question was specified by \( \pi_j \) and the probability of an incorrect answer by \( (1-\pi_j) \).

For students answering the jth question incorrectly, we specified the conditional probability (conditioned on an incorrect response) of use of the rth distractor by \( \pi_{jr} \), \( r = 1 \) to 4, with \( \sum_{r=1}^{4} \pi_{jr} = 1 \). Furthermore, for convenience, we assumed that the \( \pi_{jr} \) were ordered—that is, \( \pi_{j1} \geq \pi_{j2} \geq \pi_{j3} \geq \pi_{j4} \). Thus, the unconditional probability of use of the rth distractor was \( (1-\pi_{Correct}) \pi_{jr} \).

Using a similar notation, for the jth question we specified the observed frequency of use of a distractor as \( n_{jr} \), \( r = 1 \) to 4; therefore, the number of incorrect responses to this question was \( n_j = \sum_{r=1}^{4} n_{jr} \). The number of observed uses of the rth distractor depends on the total number of incorrect answers and the probability of using the rth distractor; the expected number of responses in which the rth distractor was used, conditional on the number of incorrect answers, is \( E(n_{jr}) = n_j \pi_{jr} \).

To identify a distractor as functioning, we defined a cutoff value, \( n_{Cutoff} \), such that if \( n_{jr} \geq n_{Cutoff} \), the distractor was classified as functioning. In our studies, we had eighty-two to eighty-eight students per class; we conducted our investigation of cutoff values using a middle number of eighty-five students. Thus, the 1 percent and 5 percent cutoff values correspond to 1 (1/85=1.18 percent) and 4 (4/85=4.71 percent) uses of a distractor in academic classes of this size. We explored these two cutoff values as well as intermediate values (use by at least two and three students). For a question with difficulties specified by chances of 90 percent, 80 percent, or 70 percent of students answering correctly, we expected to observe 8.5=9, 17, or 25.5=26 incorrect answers in eighty-five students. In our data, the number of incorrect responses to a question ranged from a minimum of one to a maximum of eighty-one, with an average of fourteen.

We utilized computer simulations to explore how the choice of cutoff value, \( n_{Cutoff} \), affected the relation between the observed and actual numbers of functioning distractors. We simulated the situation described above in which distractor use in questions answered incorrectly was distributed uniformly among the functioning distractors. Thus, for a multiple-choice question having four functioning distractors, the conditional probabilities of use of each of the distractors in a question answered incorrectly were \( \pi_{j1} = \pi_{j2} = \pi_{j3} = \pi_{j4} = \frac{1}{4} \). For a question having three functioning distractors \( \pi_{j1} = \pi_{j2} = \pi_{j3} = \frac{1}{3} \) and \( \pi_{j4} = 0 \); for a question having two functioning distractors \( \pi_{j1} = \pi_{j2} = \frac{1}{2} \) and \( \pi_{j3} = \pi_{j4} = 0 \); and for a question having one functioning distractor \( \pi_{j1} = 1 \), and \( \pi_{j2} = \pi_{j3} = \pi_{j4} = 0 \).
The computer simulations utilized the random number generator in SAS Version 9.3 (SAS Institute, Cary, NC, USA). Means of numbers of functioning distractors were calculated for 5,000 random samples for each condition of true number of functioning distractors (one to four), number of questions answered incorrectly (nine, seventeen, or twenty-six), and cutoff value ($n_{\text{cutoff}} = 1$ to 4). These choices of number of incorrect answers represented three different levels of question difficulty (90 percent, 80 percent, or 70 percent correct answers).

Figure 1 shows the mean numbers of functioning distractors determined by our simulations for nine, seventeen, and twenty-six incorrect answers to a question. For nine incorrect answers, use by at least one student to identify a functioning distractor resulted in the average observed number of functioning distractors being much closer to the true number of functioning distractors than with use of higher cutoff values; that is, the use of higher cutoff values resulted in substantial bias. This result was not surprising because cutoff values of three or four students require a minimum of twelve or sixteen incorrect answers to a question in order to have three or four functioning distractors observed. For seventeen incorrect answers to a question, substantial bias occurred for questions with four functioning distractors when cutoff values of three and four uses by students were applied. As the number of incorrect answers to a question increased to twenty-six, the criterion of use by at least one student remained the best, although there was little bias even if the cutoff value of four was used. For even more difficult questions, there would be more incorrect answers than the twenty-six illustrated, and only a small bias would occur using any of the cutoff values.

These simulations clearly showed that classifying distractors as functioning based on use by at least one student (1.17 percent of students in our case) was the best criterion where the number of incorrect answers to a question is small. Small numbers of incorrect answers could come about because of small numbers of students in a class or because students are high performing or a combination of both.

**Formula Scoring**

The formula scoring algorithm applied prospectively in the 2005-06 (SA/MC*) class was a modification of the common scoring method for multiple-choice examinations (number-correct or number-right scoring) where no point is assigned...
for an incorrect answer and +1 point (full credit) is given for a correct answer. In the multiple-choice examinations we investigated, each MCQ had five options. Therefore, the standard formula for correction for guessing consisted of awarding +1 point for a correct answer, -1/4 point for an incorrect answer, and 0 points for a question not answered. Assuming a random selection, the probability of guessing the correct answer was 1/5 (0.20) and the probability of guessing an incorrect answer was 4/5 (0.80). Therefore, using this correction for guessing, the expected value of the number of points gained due to random guessing was zero ((0.20)[1]+[0.80][-1/4]=0). In general, for \( \kappa \) possible answers per question, \(-1/(\kappa-1)\) is awarded for an incorrect answer, 0 for a question not answered, and +1 for a correct answer. This correction for guessing scoring is generally referred to as formula scoring or the standard correction for guessing. For classes other than 2005-06, formula scoring was applied retrospectively and thus had no effect on a student’s final grade nor on student behavior.

As usually applied, the parameter \( \kappa \) in the formula scoring algorithm is the number of options presented to the students. We know that students potentially could eliminate one or more distractors before making a final selection among the remaining options. Thus, the number of functioning distractors in a question, \( \delta \), can range from 0 to \( \kappa-1 \). The formula scoring algorithm based on the number of functioning distractors, \( \delta \), is awarding +1 point for a correct answer, \(-1/\delta\) points awarded for an incorrect answer, and 0 awarded for a questions left unanswered.

Each question has its own number of functioning distractors. As shown in the Appendix, calculation of the harmonic mean (the harmonic mean, \( h \), of a sample of \( n \) values of the random variable \( X \) is the reciprocal of the arithmetic mean of the \( 1/X \), that is, \( \frac{1}{h} = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{X_j} \)) of the number of functioning distractors for each question is the appropriate method to compute a single equivalent number of functioning distractors that can be applied to all questions in order to adjust each student’s multiple-choice score using formula scoring. Also (as shown in the Appendix), the harmonic mean of the equivalent numbers of functioning distractors for the students is the mathematically appropriate single value to adjust mean scores for an academic class. For example, if two questions having one and three functional distractors were both answered incorrectly by a student, then the average correction for guessing is \( \left( -\frac{1}{1} - \frac{1}{3} \right)/2 = -\frac{2}{3} = -\frac{1}{1.5} = -0.667 \), identical to subtracting the reciprocal of the harmonic mean of 1 and 3.

**Question Difficulty and Discrimination**

Difficulty of a question was assessed by the fraction of students in a class answering the question correctly: the lower the fraction answering the question correctly, the greater the difficulty of the question. Discrimination of a question was assessed by the difference in fraction (percent) of students in the upper 25 percent of the course score distribution answering the question correctly and the fraction of students in the lowest 25 percent of the course score distribution answering the question correctly.

**Statistical Analyses**

The mean numbers of functioning distractors were compared between classes using analysis of variance. The harmonic means of the number of functioning distractors also were compared between classes using analysis of variance. Standard errors of harmonic means were calculated as described by Norris. Harmonic means of the number of functioning distractors were used to calculate formula score adjustments for individual students and for classes (see Appendix). Student scores adjusted for the number of functioning distractors in the questions they answered incorrectly were compared among classes using analysis of variance. Standard errors of harmonic means were calculated as described by Norris. Harmonic means of the number of functioning distractors were used to calculate formula score adjustments for individual students and for classes (see Appendix). Student scores adjusted for the number of functioning distractors in the questions they answered incorrectly were compared among classes using analysis of variance. Average question difficulty and average question discrimination were compared among classes and numbers of functioning distractors using two-way analysis of variance. Associations of question difficulty and question discrimination with number of functioning distractors were assessed using the Spearman rank correlation coefficient.

Principal component lines, to assess the relationship between scores on the multiple-choice exams and the short-answer exams for the 2004-05 and 2005-06 classes, were estimated from the variance-covariance matrix. A first principal component line more accurately estimates the relation between scores on the multiple-choice exams and the short-answer exams than ordinary linear regression analysis because both variables are random variables. The equation of the principal component line that we used was \( Y = \beta_0 + \beta (X - 70) \) where \( Y \) is the
multiple-choice score, \( X \) is the short-answer score, \( \beta \) is the slope, and \( \beta_{00} \) is the value of \( Y \) at \( X = 70 \). The slope (\( \beta \)) of the principal component line describes the anticipated change in the multiple-choice score if the short-answer score were increased by 1 percent. The intercept parameter (\( \beta_{00} \)) is the multiple-choice score expected from the principal component line for a short-answer score of 70 percent. The value of \( Y \) for \( X = 70 \) percent is more interpretable than the usual intercept parameter that represents the value of \( Y \) for \( X = 0 \). The usual \( Y \) intercept represents an extrapolation to values that are not observed in the data, whereas the anticipated value we present is well within the range of observed data. A bootstrap procedure\(^1\) with 1,000 samples was used to estimate confidence intervals for the slope and intercept of the first principal component lines.

### Results

**Number of Functioning Distractors**

We investigated the number of functioning distractors used in three successive academic classes in an oral and maxillofacial pathology course to determine the following: 1) if differences in the examination environments were reflected in the number of functioning distractors; 2) whether the differences in number of functioning distractors paralleled the differences in student course scores between learning environments;\(^1-3\) and 3) the associations between the number of functioning distractors and question difficulty and question discrimination.

Analyzes of multiple-choice questions. Seven percent of questions (fourteen of 200) were answered correctly by all students (no functioning distractors) in the 2004-05 (SA/MC) class, 11 percent of questions (twenty-two of 200) in the 2005-06 (SA/MC*) class, and 4 percent of questions (twelve of 300) in the 2006-07 (MC) class. For those questions answered incorrectly, the relative frequencies of questions where one, two, three, or four functioning distractors, identified as being used by at least one student, in the three academic classes are shown in Figure 2. These data indicated that many questions contained one or more non-functioning distractors. The frequency distribution of number of distractors used by the 2006-07 class was significantly different (\( p \leq 0.0015 \)) from the 2004-05 and 2005-06 classes. There was no significant difference in frequency distributions between the 2004-05 and 2005-06 classes.

The average numbers of functioning distractors, defined as use by at least one, two, three, or four students, are shown in Table 2. The average number of functioning distractors in the 2006-07 (MC) class was significantly higher (\( p \leq 0.0003 \)) than the 2004-05 (SA/MC) and 2005-06 (SA/MC*) classes; there was no significant difference in the average number of functioning distractors between the 2004-05 and 2005-06 classes. As we required greater student use to define functioning distractors, the average number of functioning distractors decreased; this was anticipated based on the estimated bias (Figure 1) that occurs when a small number of questions are answered incorrectly. The intercorrelations among the estimated numbers of distractors based on using the four different cutoff values were all \( \geq 0.49 \) (Spearman rank correlation coefficients, \( p \leq 0.0001 \)).

In all three classes, the number of distractors used was inversely related to the average score for each question (Figure 3, panel A). That is, as the difficulty of a question increased, there was an increase in the number of functioning distractors. The average question scores differed significantly (\( p \leq 0.0001 \)) among the number of functioning distractors in all three classes. Also, for a particular number of distractors, the average question score was similar among the three classes as there was no significant interaction of number of functioning distractors and class (\( p = 0.8951 \), two-way analysis of variance). This inverse association of the number of functioning distractors and question difficulty also was indicated by the significant negative correlation coefficients

![Figure 2. Relative frequency of questions with functioning distractors identified by use by at least one student in questions answered incorrectly, by class](image-url)
The associations of number of functioning distractors with question difficulty and discrimination. This was accomplished by first identifying two subsets of questions having either no or one functioning distractor defined by having been used by four (5 percent) of our students. For the first question subset, where there were no functioning distractors identified by use by at least four students, the Spearman rank correlation coefficients within an academic year for number of distractors used by at least one student with question score were ≤-0.76 (p≤0.0001) and for number of distractors with question discrimination were ≥0.38 (p≤0.0009). For the second question subset, where there was one distractor identified by use by at least four students, the Spearman rank correlation coefficients within an academic year for number of distractors used by at least one student with question score were ≤-0.33 (p≤0.0052) and for number of distractors with question discrimination were ≥0.26 (p≤0.0383). These significant associations show that using a cutoff value of use by at least four students failed to identify some distractors that had the anticipated associations with question difficulty and question discrimination and thus should be considered functioning. These results support the

 between the number of distractors and the score for each question (Table 2). As the cutoff for identifying functioning distractors was increased, there was little change in the association of number of functioning distractors and question difficulty.

Questions with more functioning distractors were better able to discriminate between lower and higher performing students (Figure 3, panel B). In all three classes, the discrimination differed significantly (p≤0.0001) among the number of functioning distractors. Also, the discrimination for a particular number of functioning distractors was similar among the three classes, as there was no interaction of number of functioning distractors and class (p=0.9361, two-way analysis of variance). This direct association also was indicated by significant positive correlation coefficients between numbers of functioning distractors and question discrimination (Table 2). As the cutoff for identifying functioning distractors was increased, there was only a slight change in the association of number of distractors and question discrimination.

We then determined whether decreasing the cutoff value defining functioning distractors below the 5 percent of observed student use would affect the associations of number of functioning distractors with question difficulty and discrimination. This was accomplished by first identifying two subsets of questions having either no or one functioning distractor defined by having been used by four (5 percent) of our students. For the first question subset, where there were no functioning distractors identified by use by at least four students, the Spearman rank correlation coefficients within an academic year for number of distractors used by at least one student with question score were ≤-0.76 (p≤0.0001) and for number of distractors with question discrimination were ≥0.38 (p≤0.0009). For the second question subset, where there was one distractor identified by use by at least four students, the Spearman rank correlation coefficients within an academic year for number of distractors used by at least one student with question score were ≤-0.33 (p≤0.0052) and for number of distractors with question discrimination were ≥0.26 (p≤0.0383). These significant associations show that using a cutoff value of use by at least four students failed to identify some distractors that had the anticipated associations with question difficulty and question discrimination and thus should be considered functioning. These results support the

| Table 2. Mean number of functioning distractors for questions and Spearman rank correlations coefficients of number of functioning distractors with question score and question discrimination, by criterion for identifying functioning distractors and class |
|---|---|---|---|---|
| Criterion for Functioning Distractor | Class | Number of Questions Answered Incorrectly | Mean±SE | Number of Distractors and Question Score | Number of Distractors and Discrimination |
| | | | | | |
| Used by at least 1 student | 2004-05 (SA/MC) | 186 | 2.46±0.08 | -0.70*** | 0.59*** |
| | 2005-06 (SA/MC*) | 178 | 2.46±0.08 | -0.77*** | 0.61*** |
| | 2006-07 (MC) | 288 | 2.84±0.06*, ** | -0.74*** | 0.59*** |
| Used by at least 2 students | 2004-05 (SA/MC) | 154 | 2.13±0.08 | -0.72*** | 0.57*** |
| | 2005-06 (SA/MC*) | 140 | 2.03±0.09 | -0.74*** | 0.51*** |
| | 2006-07 (MC) | 249 | 2.41±0.07*, ** | -0.74*** | 0.52*** |
| Used by at least 3 students | 2004-05 (SA/MC) | 129 | 1.95±0.09 | -0.66*** | 0.50*** |
| | 2005-06 (SA/MC*) | 120 | 1.83±0.09 | -0.78*** | 0.44*** |
| | 2006-07 (MC) | 221 | 2.06±0.07** | -0.76*** | 0.46*** |
| Used by at least 4 students | 2004-05 (SA/MC) | 115 | 1.74±0.09 | -0.64*** | 0.52*** |
| | 2005-06 (SA/MC*) | 105 | 1.72±0.09 | -0.78*** | 0.38*** |
| | 2006-07 (MC) | 204 | 1.89±0.06 | -0.73*** | 0.41*** |

*If a question had no functioning distractors according to the specified criterion, it was not included and thus was equivalent to a question that was answered correctly by all students.

*Significantly (p≤0.05) different from 2004-05 (SA/MC)

**Significantly (p≤0.05) different from 2005-06 (SA/MC*)

***Significantly (p≤0.0001) different from zero
use by one student (1 percent of students) as a better cutoff value.

**Analyses of students.** As was the case for course scores, distractor use likely was dependent on both individual questions and individual students. The estimates of numbers of functioning distractors for each of the questions, as described in the foregoing section, were then used to calculate an average distractor use for each student in those questions they answered incorrectly. Table 3 gives the average number of distractors computed for students in each of the three classes. Regardless of the cutoff value defining a functioning distractor (one to four students), the average number of functioning distractors for the 2006-07 (MC) class was significantly higher than the average number of functioning distractors in the 2004-05 (SA/MC) and the 2005-06 (SA/MC*) classes. The average number of functioning distractors for the 2005-06 class was significantly higher than average number of functioning distractors in the 2004-05 class for cutoffs values of either one or four students. There was no significant difference between the 2004-05 and 2005-06 classes if cutoffs of two or three students were used. Thus, differences in number of functioning distractors between the 2004-05 and 2005-06 classes depend on the cutoff value used to identify a functioning distractor. Regardless of this difficulty, there is clearly no appreciable decrease in number of functioning distractors in the 2005-06 class compared to the 2004-05 class as was anticipated based on their academic performance.\(^1\),\(^3\)

The intercorrelations among the four estimates of numbers of functioning distractors for individual students (based on the four cutoff values), computed within a class, were all \( \geq 0.66 \) (Spearman rank correlation coefficients, \( p \leq 0.0001 \)). Thus, these different cutoff values for identifying functioning distractors resulted in different average levels (Table 3), while the relative relationship among the different students was preserved.

### Adjustment of Multiple-Choice Scores Using Formula Scoring

The number of functioning distractors has important ramifications for students guessing the correct answer. Obviously, a student has a greater chance of guessing the correct answer in a one distractor question (a true/false question where the chance of randomly guessing the correct answer is 1/2) than a four distractor question (chance of randomly guessing the correct answer is 1/5). In the previous section, we examined the arithmetic mean of the number of functioning distractors relative to differences in the learning environments. As we show in the Appendix, the mathematically correct value to use to adjust the multiple-choice scores using formula scoring is not the commonly used arithmetic mean but instead is the less-well-known weighted harmonic mean of the numbers of functioning distractors. The harmonic means for the three classes are shown in Table 3. The harmonic means show the same pattern among the different classes as reported in the previous section for arithmetic average number of functioning distractors.

Inspection of the harmonic means (Table 3) indicated considerable variation in the estimates of the required correction (the amount subtracted for an incorrect answer) due to choice of cutoff value used to define a functioning distractor. To help select among these different \( \delta \) values for adjusting multiple-
The values of $\delta=2.52$ for the 2004-05 class and $\delta=3.04$ for the 2005-06 class resulted in principal component lines that were parallel (slope=1.0) to the line of equality.

Figure 4 shows the relationship between adjusted multiple-choice scores for students and their respective short-answer scores. Multiple-choice scores were adjusted in two ways: 1) adjusted multiple-choice scores were calculated for each student using the $\delta$ for each question determined by the counting of distractors used by at least one student; and 2) adjusted multiple-choice scores were calculated using a constant value in a class for all questions, namely, the harmonic means of $\delta=2.68$ for the 2004-05 (SA/MC) class and $\delta=2.81$ for the 2005-06 (SA/MC*) class. The adjustments using the harmonic mean number of functioning distractors for a class were markedly similar to the adjustment using a separate $\delta$ for each question and student; this strongly supports using a constant $\delta$ for adjustment and eliminates the need for more complicated adjustment schemes. The equations (slopes and predicted multiple-choice score for a short-answer score of 70 percent) of the principal component lines for these choices of $\delta$ are shown in Table 4.

### Table 4. Arithmetic and harmonic mean number of functioning distractors for students, by criterion for identifying functioning distractors and class

<table>
<thead>
<tr>
<th>Criterion for Functioning Distractor</th>
<th>Class</th>
<th>Number of Students</th>
<th>Arithmetic Means Mean±SE</th>
<th>Harmonic Means Mean±SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used by at least 1 student</td>
<td>2004-05 (SA/MC)</td>
<td>88</td>
<td>3.07±0.02</td>
<td>2.68±0.03</td>
</tr>
<tr>
<td></td>
<td>2005-06 (SA/MC*)</td>
<td>82</td>
<td>3.26±0.02*</td>
<td>2.81±0.04*</td>
</tr>
<tr>
<td></td>
<td>2006-07 (MC)</td>
<td>88</td>
<td>3.47±0.02* **</td>
<td>3.14±0.02* **</td>
</tr>
<tr>
<td>Used by at least 2 students</td>
<td>2004-05 (SA/MC)</td>
<td>88</td>
<td>2.73±0.02</td>
<td>2.23±0.02</td>
</tr>
<tr>
<td></td>
<td>2005-06 (SA/MC*)</td>
<td>82</td>
<td>2.68±0.02</td>
<td>2.11±0.02*</td>
</tr>
<tr>
<td></td>
<td>2006-07 (MC)</td>
<td>88</td>
<td>2.99±0.02* **</td>
<td>2.51±0.02* **</td>
</tr>
<tr>
<td>Used by at least 3 students</td>
<td>2004-05 (SA/MC)</td>
<td>88</td>
<td>2.43±0.02</td>
<td>1.89±0.01</td>
</tr>
<tr>
<td></td>
<td>2005-06 (SA/MC*)</td>
<td>82</td>
<td>2.42±0.02</td>
<td>1.89±0.02</td>
</tr>
<tr>
<td></td>
<td>2006-07 (MC)</td>
<td>88</td>
<td>2.59±0.02* **</td>
<td>2.09±0.01* **</td>
</tr>
<tr>
<td>Used by at least 4 students</td>
<td>2004-05 (SA/MC)</td>
<td>88</td>
<td>2.16±0.02</td>
<td>1.68±0.01</td>
</tr>
<tr>
<td></td>
<td>2005-06 (SA/MC*)</td>
<td>82</td>
<td>2.24±0.02*</td>
<td>1.74±0.01*</td>
</tr>
<tr>
<td></td>
<td>2006-07 (MC)</td>
<td>88</td>
<td>2.35±0.02* **</td>
<td>1.86±0.01* **</td>
</tr>
</tbody>
</table>

*Significantly (p≤0.05) different from 2004-05 (SA/MC)
**Significantly (p≤0.05) different from 2005-06 (SA/MC*)

choice scores, we compared the adjusted scores for individual students on the multiple-choice portions of the examinations with their respective scores on the short-answer portions of the examinations in the 2004-05 (SA/MC) and the 2005-06 (SA/MC*) classes. The equations of the principal component lines relating the adjusted multiple-choice scores with short-answer scores are shown in Table 4 for the four different cutoff values for identifying a functioning distractor. These adjusted multiple-choice scores were calculated for each student with $\delta$ values for each question determined by the counting of distractors used by at least one, two, three, or four students. As the cutoff value increased, the slope of the principal component line increased and the predicted multiple-choice scores decreased for a short-answer score of 70 percent. The lines describing the adjustment with $\delta$ for each question calculated by counting the distractors used by at least one student were closer to parallel with the line of equality than were the lines derived from the calculation of $\delta$ using higher cutoff values. Thus, the number of functional distractors determined using this cutoff of use by one student resulted in similar scales for the two examination question formats. Having similar scales implies that a change of 1 percent in short-answer score was associated with a change of close to 1 percent in the multiple-choice score. Additionally, we determined the precise value of $\delta$ that would result in the principal components line relating adjusted multiple-choice scores to short-answer scores having a slope of 1.0.

Choosing a cutoff value for identifying a functioning distractor is a subjective decision that can affect the adjustment of scores. A higher cutoff value may result in a smaller number of distractors being counted as functioning, potentially leading to a more lenient adjustment of scores. Conversely, a lower cutoff value may result in a larger number of distractors being counted as functioning, potentially leading to a more strict adjustment of scores. The choice of cutoff value should be based on the specific context and goals of the examination. It is important to consider the potential impact of the cutoff value on the interpretation of the results and the fairness of the examination.
as reported previously have slopes less than 1.0 and indicate inflation in multiple-choice relative to short-answer scores.

Table 5 shows the average uncorrected and corrected scores for the three classes; the corrected scores were calculated using several values of δ. The use of the harmonic mean number of functioning distractors in the formula scoring algorithm represented the best estimate of the average actual knowledge levels of these three classes based on the multiple-choice examinations. We previously used correction for random guessing among five options (δ=4), which turned out to be an underestimate resulting in score inflation. Bringing together these different δ values into an integer value of δ=3 for all three classes achieved similar averages for them (differences ≤0.6%) to the class averages of scores adjusted using the δ’s estimated by the harmonic means. For all the adjustments shown in Table 5, the average adjusted scores in the 2004-05 (SA/MC) and 2005-06 (SA/MC*) classes were significantly different from the 2006-07 (MC) class. The adjusted averages for the 2004-05 and 2005-06 classes were not significantly different for any adjustment.

Table 4. Slope and intercept for equations of straight lines for first principal components between adjusted multiple-choice and short-answer scores, by criterion for estimation of formula scoring parameter and class

<table>
<thead>
<tr>
<th>Class</th>
<th>Criterion for Formula Scoring Adjustment</th>
<th>δ</th>
<th>Estimated β (95% CI)</th>
<th>Estimated β&lt;sub&gt;unc&lt;/sub&gt; (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004-05 (SA/MC)</td>
<td>Used by at least 1 student</td>
<td>Variable&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.99 (0.89-1.12)</td>
<td>69.69 (64.76-69.01)</td>
</tr>
<tr>
<td></td>
<td>Used by at least 2 students</td>
<td>Variable&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.06 (0.95-1.20)</td>
<td>65.12 (62.72-67.24)</td>
</tr>
<tr>
<td></td>
<td>Used by at least 3 students</td>
<td>Variable&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.10 (0.98-1.24)</td>
<td>63.79 (61.38-65.90)</td>
</tr>
<tr>
<td></td>
<td>Used by at least 4 students</td>
<td>Variable&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.13 (1.01-1.27)</td>
<td>62.63 (60.15-64.78)</td>
</tr>
<tr>
<td></td>
<td>Parallel with line of equality</td>
<td>2.52</td>
<td>1.00 (0.89-1.13)</td>
<td>66.56 (64.28-68.54)</td>
</tr>
<tr>
<td></td>
<td>Harmonic mean&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.68</td>
<td>0.98 (0.87-1.10)</td>
<td>67.18 (64.95-69.14)</td>
</tr>
<tr>
<td></td>
<td>Uncorrected</td>
<td>∞</td>
<td>0.66 (0.58-0.74)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>76.77 (75.25-78.22)</td>
</tr>
<tr>
<td></td>
<td>Random guessing among 5 items</td>
<td>4</td>
<td>0.87 (0.77-0.98)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>70.40 (68.40-72.18)</td>
</tr>
<tr>
<td>2005-06 (SA/MC*)</td>
<td>Used by at least 1 student</td>
<td>Variable&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.08 (0.81-1.42)</td>
<td>66.19 (59.98-70.76)</td>
</tr>
<tr>
<td></td>
<td>Used by at least 2 students</td>
<td>Variable&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.21 (0.91-1.61)</td>
<td>62.91 (55.57-68.14)</td>
</tr>
<tr>
<td></td>
<td>Used by at least 3 students</td>
<td>Variable&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.28 (0.95-1.69)</td>
<td>61.50 (53.83-67.11)</td>
</tr>
<tr>
<td></td>
<td>Used by at least 4 students</td>
<td>Variable&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.29 (0.96-1.71)</td>
<td>61.05 (53.34-66.59)</td>
</tr>
<tr>
<td></td>
<td>Parallel with line of equality</td>
<td>3.04</td>
<td>1.00 (0.76-1.31)</td>
<td>67.67 (62.15-71.69)</td>
</tr>
<tr>
<td></td>
<td>Harmonic mean&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.81</td>
<td>1.03 (0.79-1.35)</td>
<td>66.69 (61.23-71.03)</td>
</tr>
<tr>
<td></td>
<td>Uncorrected</td>
<td>∞</td>
<td>0.63 (0.49-0.81)</td>
<td>74.76 (74.25-79.98)</td>
</tr>
<tr>
<td></td>
<td>Random guessing among 5 items</td>
<td>4</td>
<td>0.93 (0.71-1.21)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>69.68 (64.70-73.50)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Correction calculated for each student using δ for each question.
<sup>b</sup>Table 3 with functioning distractors identified by use by at least one student.
<sup>c</sup>Blank awarded +0.20 points, 2005-06 (SA/MC*) class only.

Discussion

Identifying functioning distractors requires setting a frequency of use equal to or greater than a defined cutoff value. Haladyna and Downing required that a functioning distractor had to have an observed frequency of use equal to or greater than 5 percent of students. More recently, Rogausch et al. suggested 1 percent of students as an alternative cutoff value because they believed that distractors could not generally be expected to attract more than 1 percent of the well-prepared examinees taking a high-stakes medical examination. The ability to accurately estimate the number of functioning distractors depends critically on the number of incorrect answers to a question (Figure 1). Small numbers of incorrect answers may occur as a result of small class size, because students are high performing, or a combination of the two; the latter is generally the situation in graduate level courses in dental and medical schools. In the relatively small classes (eighty-two to eighty-eight students) of high-performing dental students that we studied (median course score about 85 percent), our results supported the use of the 1
than the 2006-07 (MC) class (Table 3). This finding was anticipated in view of the significantly lower overall course performance by the 2006-07 class. However, it was not anticipated that the number of functioning distractors would be lower in the 2004-05 class than in the 2005-06 class where course scores among the lower performing students were significantly higher.

Thus, differences in numbers of functioning distractors do not always correspond to differences in course performance and likely reflect other aspects of the learning environment. Of note, similar conclusions regarding differences among the several examination environments were obtained using the four different cutoff values to estimate number of functioning distractors (Tables 2 and 3).

All cutoff values for identifying functioning distractors (one to four students) resulted in estimates of numbers of functioning distractors in which higher numbers of functioning distractors were associated with percent criterion rather than higher cutoff fractions, in accord with the findings of Rogausch et al. In classes of the size we studied, the 1 percent criterion corresponds to use by one student. This might raise concern that the use of a distractor might have been made by mistake; however, a cutoff value of use by two students did not perform as well as use by one student (Figure 1, Table 4).

Two aspects of the learning environment—examination question format and scoring procedure—affected the average number of functioning distractors as they did in the previously reported overall student performance (course score). The highest number of functioning distractors occurred in the 2006-07 (MC) class examined using all multiple-choice questions. The numbers of functioning distractors in the 2004-05 (SA/MC) and 2005-06 (SA/MC*) classes, given half multiple-choice and half short-answer questions, were significantly lower than the 2006-07 (MC) class (Table 3). This finding was anticipated in view of the significantly lower overall course performance by the 2006-07 class. However, it was not anticipated that the number of functioning distractors would be lower in the 2004-05 class than in the 2005-06 class where course scores among the lower performing students were significantly higher. Thus, differences in numbers of functioning distractors do not always correspond to differences in course performance and likely reflect other aspects of the learning environment. Of note, similar conclusions regarding differences among the several examination environments were obtained using the four different cutoff values to estimate number of functioning distractors (Tables 2 and 3).

Figure 4. Scatter diagrams and principal component lines relating adjusted scores on multiple-choice exams to same students’ scores on short-answer exams

Note: Panel A shows academic year 2004-05 (SA/MC), and Panel B shows academic year 2005-06 (SA/MC*). In Panel A, filled circles represent multiple-choice scores adjusted for each student and δ for each question calculated by counting of functioning distractors in each question; the solid line represents the first principal component of these data. The open circles represent multiple-choice scores corrected using δ=2.68; the long dashed line represents the line of equality. In Panel B, filled circles represent multiple-choice scores adjusted for each student and δ for each question calculated by counting of functioning distractors in each question; the solid line represents the first principal component of these data. The open circles represent multiple-choice scores corrected using δ=2.81; the long dashed line represents the first principal component. The short dashed line represents the line of equality.
As usually applied, the formula scoring algorithm to correct for guessing awards +1 point for a correct answer, $-\frac{1}{(\kappa - 1)}$ points for an incorrect answer, and 0 points for a question unanswered; $\kappa$ is the number of options (the correct answer, functioning and non-functioning distractors) in the question. Using only the number of functioning options (the correct answer and $\delta$, the number of functioning distractors) would represent a more realistic adjustment for guessing; thus, $-\frac{1}{\delta}$ should be used for an incorrect answer. Importantly, our current studies demonstrate that use of a constant value of $\delta$ for a class produces adjustments similar to adjustments based on the $\delta_j$ for each question and the questions that individual students answered incorrectly (Figure 4). The mathematically correct value of $\delta$ to be used in formula scoring of multiple-choice examinations for a class is the harmonic mean of the number of functioning distractors (see Appendix).

Because of the limited number of questions answered incorrectly, applying different cutoff values for identifying functioning distractors resulted in varying estimates of the number of functioning distractors to be used in the formula scoring algorithm for a class (Table 3). For the 2004-05 (SA/MC) and the 2005-06 (SA/MC*) classes, we investigated which of the $\delta$ values (harmonic means, Table 3) resulted in the first principal component line relating adjusted multiple-choice scores to short-answer scores being close to parallel with the line of equality (slope=1.0). We targeted this slope of 1.0 because an incremental change of 1 percent in score on MCQs is equivalent to an incremental change of 1 percent in score on short-answer questions (that is, the scales are equivalent). Calculations of $\delta$ based on identifying a functioning distractor by use by at least one student resulted in slopes of the first principal component lines being much closer to 1.0 (Table 4).

Table 5. Unadjusted and retrospectively adjusted multiple-choice scores, by formula scoring adjustment parameter and class

<table>
<thead>
<tr>
<th>Class</th>
<th>Unadjusted</th>
<th>$\delta=4$</th>
<th>$\delta=3$</th>
<th>Harmonic Mean $\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean±SE</td>
<td>Mean±SE</td>
<td>Mean±SE</td>
<td>$\delta$</td>
</tr>
<tr>
<td>2004-05 (SA/MC)</td>
<td>85.2±0.7*</td>
<td>81.6±0.8*</td>
<td>80.3±0.9*</td>
<td>2.68</td>
</tr>
<tr>
<td>2005-06 (SA/MC*)</td>
<td>86.5±0.5*</td>
<td>83.4±0.7*</td>
<td>82.3±0.7*</td>
<td>2.81</td>
</tr>
<tr>
<td>2006-07 (MC)</td>
<td>82.8±0.7</td>
<td>78.5±0.9</td>
<td>77.1±0.9</td>
<td>3.13</td>
</tr>
</tbody>
</table>

*Significantly different from 2006-07 (MC). For unadjusted score, blank awarded the expected gain due to random guessing of +0.20 points, 2005-06 (SA/MC*) class only.
Thus, all our findings support the criterion that use of a distractor by at least one student (1 percent) to identify functioning distractors provided for a better estimate of number of functioning distractors in our educational environment. This is because our class sizes were small (about 100 students) and the students were high performing, both of which resulted in a relatively small number of students answering a question incorrectly. The best cutoff value for use in other courses with larger class sizes and/or less well-prepared students where more questions are answered incorrectly would have to be determined for each situation.

Use of the harmonic mean number of functioning distractors in the formula scoring algorithm represents a substantial improvement in reducing the score inflation of multiple-choice scores over formula scoring using simply the number of options presented to the students in the exam questions. Nonetheless, our model still assumes that a student who does not know the correct answer randomly guesses among a reduced subset of options consisting of the correct answer and the remaining incorrect functional distractors. Thus, a $\delta$ less than the four presented incorrect options is consistent with the interpretation that students were actually ruling out one or two incorrect distractors before answering a question.

The $\delta$ for a class also was estimated by forcing the first principal component line relating multiple-choice scores and short-answer scores to be parallel to the line of equality and represented an empirical fitting of a transformation that makes the scales of the two examination question formats the same. This estimate of $\delta$ did not depend on a model of guessing; the only connection with the formula scoring algorithm is that the same scoring procedure (+1 for a correct answer, $-1/\delta$ for an incorrect answer, 0 unanswered) was used. On the other hand, the use of the harmonic mean in the formula scoring algorithm depended on a random guessing model. For a given class, the similarity (Table 4) of the two values of $\delta$ obtained by these two completely different approaches provided validation for use of the formula scoring algorithm with the harmonic mean of the numbers of functional distractors used to determine the points subtracted for an incorrect answer.

It has been suggested that instructors should “develop as many functional distractors as feasible.”13 Our results confirm that as the number of functioning distractors increased, question difficulty and question discrimination both increased. Our analysis also indicates that the questions used in the examinations in the oral and maxillofacial pathology course did not always achieve the targeted number of four functioning distractors for each question. If faculty members cannot develop the target number of functioning distractors, they might consider reducing the number of distractors presented to the students; moreover, the number of options need not be the same for all questions. Rather than devoting more time in trying to produce greater numbers of distractors, instructors might make more efficient use of their time by developing questions that are more challenging and that better estimate students’ critical thinking skills. If students have to read fewer items, they would have more time to address more questions; this would be expected to improve examination reliability. Haladyna et al.14 argued that “three options are sufficient in most instances” and “the effort of developing the fourth option (the third plausible distractor) is probably not worth it.” A meta-analysis by Rodriguez15 led to his statement that only three distractors are “feasible.” Furthermore, Rodriguez16 also argued that three items (correct answer and two distractors) are optimal. We strongly believe that faculty members should routinely investigate distractor use as part of their assessment of exam characteristics in addition to question difficulty and discrimination. Unused or infrequently used distractors should be eliminated or modified for use in future exams. As Haladyna and Downing16 stated, “The key in distractor development is not the number of distractors but the quality of distractors.”

However, while these recommendations seem laudable, reducing the number of distractors increases the likelihood of students’ guessing the correct answer; this mandates formula scoring to adjust for the score inflation of multiple-choice examinations. For example, if a student knows 60 percent of the information covered by a 100-question examination, he or she could potentially answer forty questions incorrectly. If there were five options per question, a student would be expected to randomly guess (no knowledge) eight correct answers of the forty questions, resulting in an examination score of 68 percent. If there were four, three, or two options per question, a student would be expected to randomly guess ten, thirteen and a third, or twenty correct answers of the forty, yielding scores of 70 percent, 73.3 percent, or 80 percent, respectively. The difference in expected score between the five-option and four-option questions does not appear great, even for this poorly performing student; however, even this difference in number of options would affect a
student’s ability to achieve a score equal to or greater than a common passing score of 70 percent. Thus, in all these situations, including random guessing among five options, the inflation in score above the 60 percent true knowledge level is somewhat worrisome. Table 5 further emphasizes the inflation of multiple-choice scores by showing the effects of various choices of $\delta$ on the adjusted scores in the three classes we investigated.

We believe that faculty members should routinely apply formula scoring retrospectively to obtain a better estimate of the actual levels of knowledge of their students. As we have reported previously, retrospective formula scoring for random guessing in five-option multiple-choice questions ($\delta=4$) significantly improved the validity of multiple-choice examination scores relative to short-answer examination scores. The current study shows additional improvements in removing score inflation in multiple-choice examinations by determining the number of functioning distractors using an appropriate cutoff value and utilizing the harmonic mean number of functioning distractors in the formula scoring algorithm (Table 5). For faculty members dealing with relatively small classes such as ours, we recommend that a functioning distractor should be defined as one that is used by at least one student. This approach is operationally simple and can easily be implemented with minimal programming.

While retrospective use of formula scoring would provide valuable information to better estimate student academic achievement, it would have no influence on student learning behaviors. The calculation of the harmonic mean number of functioning distractors can only be accomplished retrospectively. Formula scoring applied prospectively significantly enhanced student achievements; however, the adjustment parameter has to be preselected so that students can be informed of the precise scoring procedure. As shown in Table 5, in our environment a value such as $\delta=3$ could be utilized as a “standard” value, so that formula scoring could be used prospectively to increase validity of multiple-choice examination scores and also to enhance student academic performance. However, because we have shown that distractor use is affected by several aspects of the learning environment, estimation of distractor utilization and calculation of the adjustment parameter would have to be determined for each individual course and would have to be determined whenever there were appreciable changes in the learning environment within a course.

## Conclusion

In an oral and maxillofacial pathology course, the number of functioning distractors in multiple-choice questions used by second-year dental students was significantly influenced by examination question format and scoring procedure. Identifying functioning distractors through use by at least 1 percent of students as a cutoff value performed better than use of higher cutoff values. As anticipated, greater numbers of functioning distractors were associated with higher question difficulty and discrimination. Dental students typically ruled out one or more non-functioning distractors in multiple-choice examinations comprised of five-option questions. This distractor exclusion exacerbated examination score inflation by fostering more effective guessing of the correct answer, thereby overestimating student academic achievements. This study strongly supports evaluating distractor use in addition to evaluating question difficulty and discrimination, as part of the customary analyses of examinations. Furthermore, evaluation of distractor use should include calculation of the harmonic mean number of functioning distractors to be used in the retrospective application of the formula scoring algorithm to more accurately assess student academic achievement.

## REFERENCES

Mathematical Calculations

We consider an academic class consisting of n students examined using q multiple-choice format questions. The jth (j=1,2,\ldots,q) question has \(k_j\) options and thus, \(d_j=k_j-1\) distractors. The number of items does not have to be the same for all questions.

The ith (i=1,2,\ldots,n) student answers \(r_i\) questions correctly, \(w_i\) questions incorrectly, and leaves \(b_i\) questions unanswered (\(r_i+w_i+b_i=q\)). One point is awarded for a correct answer, 0 points are awarded for a question left unanswered, and wrong answers are penalized by a correction for random guessing using the \(d_j\) for the missed questions. The number of distractors for an individual question, \(d_j\), is the same for all students. Note that, in this appendix, we have not included the usual multiplication by 100 so that a score would be presented as a percent.

The formula score, \(S_i\), for ith student, computed from the number of questions answered correctly, the number of questions answered incorrectly, and the number of questions left unanswered, is

\[
S_i = \frac{1}{q} \left( r_i - \sum_{j=1}^{q} \frac{1}{d_j} \right).
\]

Straightforward algebra yields

\[
S_i = \frac{1}{q} \left( r_i - w_i \left( \frac{1}{\frac{1}{w_i}} + \frac{1}{\sum_{j=1}^{q-1} \frac{1}{d_j}} \right) \right)
\]

\[
S_i = \frac{1}{q} \left( r_i - w_i \frac{1}{h_i} \right)
\]

where \(\frac{1}{h_i} = \frac{1}{w_i} \sum_{j=1}^{q-1} \frac{1}{d_j}\) is the reciprocal of the harmonic mean of the \(d_i\) computed for the set of questions missed by the ith student. This enables us to define an equivalent number of distractors (that is, use of this number in the formula scoring algorithm instead of the individual \(d_i\) results in the same formula score for this student) used by this student as \(\delta_{\text{Student } i} = h_i\).

The formula score for the academic class, \(\bar{S}\), is obtained as the average of the formula score for the students, that is,

\[
\bar{S} = \frac{1}{n} \sum_{i=1}^{n} S_i = \frac{1}{n q} \sum_{j=1}^{q} \left( r_j - w_j \frac{1}{h_j} \right).
\]

The total number of questions missed is \(m = \sum_{i=1}^{n} w_i\). Again, straightforward algebra yields

\[
\bar{S} = \frac{1}{n q} \sum_{i=1}^{n} (r_i) - \frac{1}{n q} \sum_{i=1}^{n} w_i \frac{1}{h_i},
\]

\[
\bar{S} = \frac{1}{n q} \sum_{i=1}^{n} (r_i) - \frac{m}{n q} \left( \frac{1}{H} \right)
\]

\[
\bar{S} = \frac{1}{n q} \sum_{i=1}^{n} (r_i) - \frac{m}{n q} \frac{1}{H}
\]

where \(\frac{1}{H} = \frac{1}{m} \sum_{i=1}^{n} w_i \frac{1}{h_i} = \frac{1}{m} \sum_{i=1}^{n} w_i \left( \frac{1}{w_i} \sum_{j=1}^{q-1} \frac{1}{d_j} \right) = \frac{1}{m} \sum_{i=1}^{n} \left( \frac{1}{d_{\text{Student } i}} \right)\) is the reciprocal of the harmonic mean over all questions answered incorrectly. We also note that this is a weighted harmonic mean of the equivalent number of distractors for the individual students, that is, \(\frac{1}{H} = \frac{1}{m} \sum_{i=1}^{n} \frac{1}{d_{\text{Student } i}}\).

The harmonic mean, \(H\), represents an equivalent number of distractors that is appropriate for correcting a class average. That is, \(\delta_{\text{Class}} = H\) and \(\bar{S} = \frac{1}{n q} \sum_{i=1}^{n} (r_i) - \frac{m}{n q} \frac{1}{\delta_{\text{Class}}}.\)