Correcting for Guessing on Multiple-Choice Exams

Dear Dr. Alvares:

Thomas J. Prihoda et al. published an article concerning correcting for guessing on multiple-choice examinations in the April 2006 issue of the Journal of Dental Education. They correlated scores on four tests in a multiple-choice format (without correction for guessing) with scores on four short-answer tests given at the same time. A post hoc analysis was performed on the same data (this time applying the standard formula giving a penalty for guessing). The authors concluded, as expressed in their title, that “Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course.”

There are six concerns that make this conclusion untenable.

First, validity is a characteristic of decisions made from data, not a property of scores in the abstract. The authors appear to be making an argument for concurrent validity: the short-answer test is the “gold standard,” and measures that correlated more highly with it are more valid. If so, we might be better advised to skip the multiple-choice format altogether rather than trying to patch it up. In the discussion section of the article, Frederick Lord is cited as a model for this research. He argued that reliability is expected to increase on theoretical grounds under circumstances resembling this study. Lord correctly identifies the issue as “reliability,” not test validity.

Second, the correction for guessing was applied post hoc, and students were not told to respond based on an expectation that a penalty would be applied. It is reasonable to believe that this information, had it been available, would have influenced the answering strategies of at least some students. It would be unwise to generalize from the post hoc analysis to a practice of correction for guessing in which students are informed in advance.

Third, the authors’ claim that the correlation between corrected multiple-choice and short-answer scores is greater than the correlation between uncorrected multiple-choice and the same short-answer scores cannot be verified based on available data. A simulation method (bootstrapping) was used by the authors to estimate the parameters for the statistical test of this hypothesis. Parameters (unspecified) of the bootstrap model are critical because of lack of independence among the four examinations (across the same students) and lack of independence among the corrected and uncorrected multiple-choice tests (across the same criterion variable). Two other statisticians and I reconstructed reasonable tests based on the confidence intervals provided in the Prihoda et al. article. We were able to approximate the reported test value (p=.015) only when using a sample size of 352. This is four times the number of students completing the course, and the confusion may result from using the total of all tests for all students (n=352) instead of the average of the four test scores (n=88).

Fourth, there are more appropriate adjustments available than the correction for guessing procedure. The slopes and intercepts of principal components fittings of the relationship between multiple-choice and short-answer tests, as reported in Table 3 in the article, show a clear and verifiable advantage for the correction for guessing process. In particular, this adjustment deflates the high scores of weak students. The same sort of “tidying up” could also be accomplished by any of several transformations on the data, without resorting to the correction for guessing formula. The easiest and most defensible such adjustment is to make higher cut scores for the low grades. Most grading scales are arbitrary (the authors report a traditional 90 percent, 80 percent, 70 percent, 60 percent system), without offering any justification that these cuts correspond to meaningful discipline or educational characteristics or that the intervals should be equal in width.

Fifth, the correction for guessing formula is almost never used. Despite its having been known and theoretically evaluated for decades, there is no high-stakes standardized testing program (such as DAT, NBDE, or their counterparts) that employs the system. The authors mention none in their article. It is easier to add a few questions (automatically increasing the test reliability thereby) or to adjust
Sixth, the authors have not made it easy to validate the accuracy of their research. When my preliminary checks of the reported data could not be reconciled with their article, I consulted two colleagues. That failing, I corresponded with one of the authors in hopes of getting the dataset itself because the analysis used is very sophisticated. That request was declined. This puts me in mind of instructors who refuse to return exams to students on the grounds that they are thus keeping the tests secure. Full disclosure promotes learning and is the only way to protect against “unauthorized release” to a selected set of students. It may come as a surprise, but a partially released test is expected to have a higher reliability than one that is fully available. The few who have advanced knowledge will perform consistently well. A high correlation does not automatically equate with high validity.

—David W. Chambers, Ed.M., M.B.A., Ph.D.
Associate Dean for Academic Affairs and Scholarship
Arthur A. Dugoni School of Dentistry
University of the Pacific
2155 Webster Street
San Francisco, CA 94115

REFERENCES

The authors respond:

The authors appreciate the opportunity to respond to David W. Chambers. Nothing in his letter casts any doubt on our conclusion and certainly does not make our conclusion untenable. His statements of concern and his terminology are imprecise, and he attributes conclusions to us that are not in the article. Further, many of his remarks were already addressed in our article.

Our study was based on data from an oral and maxillofacial pathology course in which the same subject matter was tested in the same examination using both short-answer format questions and multiple-choice format questions. We demonstrated that the standard correction for guessing applied to multiple-choice scores resulted in better agreement with the scores on the short-answer examinations; this improved agreement was shown visually, by principal components analysis, and by use of intraclass correlation coefficients. Our results strongly supported the use of correction for guessing to obtain a more valid score on multiple-choice format examinations.

In the opening paragraph of his letter, Chambers states that we “correlated scores.” The use of the term “correlation” without qualification is imprecise and potentially misleading. As stated above and in the article, we reported intraclass correlation coefficients. If Chambers instead meant the more commonly used Pearson correlation coefficient, this statistic was not used in our article as it would not have been an appropriate statistic (see the article, page 379, right-hand column, lines 37-39). The Pearson correlation coefficient measures linear association, whereas we measured agreement between the scores on examinations using a short-answer format and scores on examinations using a multiple-choice format. As noted in our article, the reference by Bland and Altman makes the point about agreement clear. As we pointed out (page 379, right-hand column, lines 28-32), the Pearson correlation coefficient for uncorrected multiple-choice examination scores with short-answer scores is identical to the Pearson correlation coefficient for retrospectively corrected multiple-choice scores with short-answer scores. This is because, retrospectively, the correction is a linear transformation of the uncorrected multiple-choice score and such a linear transformation does not change the Pearson correlation coefficients.

The following are our responses to Chambers’s six criticisms. First, his issue apparently is with our use of the term “validity.” Our study investigated construct validity described by Cronbach (Table 14.1) as...
addressing the question “Does the test measure the attribute it is said to measure?” The criteria in our article for measuring the attribute of interest were the scores on the short-answer format examinations. We clearly demonstrated significantly improved agreement resulting from applying a correction for guessing to the multiple-choice scores. We did not claim the short-answer examination was a “gold standard.” We stated that it “eliminates or at least greatly reduces the potential for guessing the correct answer” (page 378, left-hand column, lines 12-13) and reiterated this point in the Discussion section (page 382, right-hand column, lines 8-13). We do believe that there are better means to measure a student’s knowledge than the use of multiple-choice format examinations and using better examinations is preferred to “patching-up” the multiple-choice format examinations. However, multiple-choice examinations are in wide use, and our results strongly support correction for guessing as a means for improving the match with scores from short-answer format examinations.

The Lord reference was used correctly in our article, both as a source for the standard correction for guessing (page 379, right-hand column, line 25) and to discuss the anticipated effects of a prospective application of correction for guessing on reliability (page 384, right-hand column, lines 11-13).

Second, Chambers criticizes the fact that a retrospective correction for guessing was investigated. In our article, the retrospective application of the standard correction for guessing was clearly described. Our retrospective study clearly demonstrated that guessing is occurring in multiple-choice examinations at the expected level, that grades are inflated as a result of this guessing, and that this inflation may potentially be corrected by simply implementing the correction for guessing. This demonstration is only possible in a retrospective application and represents a strength of the study. We clearly stated in the Discussion that we had not studied correction for guessing after students were told such a correction would be applied (page 384, right-hand column, lines 16-21).

Third, Chambers misstates our “claim” and goes on to state that “The authors’ claim . . . cannot be verified based on available data.” Our conclusion was supported by three different methods: a) qualitatively (visually)—Figure 1 showed much greater agreement of corrected multiple-choice and short-answer scores; b) quantitatively—the principal components analyses showed better aggregate agreement after correction for guessing; and 3) quantitatively—intraclass correlation coefficients showed better agreement for individual students after correction for guessing.

Once again, in his description of our “claim,” Chambers raises the issue of the use of correlation. As pointed out in the response to the opening paragraph, this is imprecise and likely misleading. Chambers also is vague with respect to parameters of the bootstrap that he thinks need to have been specified. The lack of independence among the four examinations was properly handled in all our analyses. We used mixed models to estimate reliability; the intraclass correlation coefficients measuring agreement were presented separately for each of the four examinations; the average of the four tests was used in the principal components analysis (n=88) and in the intraclass correlation coefficients for agreement of averages. Chambers correctly noted the lack of independence between uncorrected and corrected scores. He apparently is not aware that this lack of independence was correctly included in the bootstrap analysis, which is a strength of that analysis. Another strength is that the bootstrap does not require the test scores to be normally distributed for its result to be valid. Although Chambers and the two unnamed statisticians were aware of this non-independence, they apparently ignored this concern to make an analysis in which they assumed independence. They then used their incorrect analysis to criticize our correct analysis.

Fourth, Chambers states, “There are more appropriate adjustments available than the correction for guessing procedure,” but does not give any specific recommendations. The standard correction for guessing is simple to implement and, as shown in our article, resulted in substantially better agreement with the scores from the short-answer examinations. This better agreement was indicated qualitatively (visually) and quantitatively by the principal component lines. In fact, Chambers states, “The slopes and intercepts of principal components fittings of the relationship between multiple-choice and short-answer tests, as reported in Table 3, show a clear and verifiable advantage for the correction for guessing process.” Furthermore, the agreement for individual students was improved by correction for guessing as indicated by the intraclass correlation coefficients (page 381, Table 2).

Our article (page 383, Figure 2 and page 383, left-hand column, lines 14-17) describes the grade inflation due to guessing on multiple-choice examinations and shows how grade cutoff points could be changed to deal with average grade inflation. The
grade cutoff points of 60, 70, 80, and 90 percent were clearly described as common (page 383, footnote to Figure 2) and were used only as an example for discussion; there was no recommendation.

Fifth, Chambers uses the lack of widespread use of correction for guessing as a criticism of our article. The issue is not what others have chosen to do. Our results strongly support using correction for guessing to obtain a more valid score (see response to opening paragraph). Importantly, our article was based on data from a course in which the same subject matter was tested at the same examination using the two different formats. Our investigation within this unique situation represents an important contribution to the literature on correction for guessing.

Chambers states, “It is easier to add a few questions (automatically increasing the test reliability thereby) or to adjust the cut scores at the low end of the distribution or, better, to set the cut scores rationally (to increase validity).” As we pointed out in the article, additional questions provide the potential to increase reliability (page 384, left-hand column, lines 22-25). Also as pointed out (response to item 4), changing grade cutoff points could be used to deal with average grade inflation due to guessing on multiple-choice examinations. However, we do not support this approach to deal with grade inflation (see final paragraph of our response).

Sixth, Chambers describes our analysis as “sophisticated.” We believe that a “sophisticated” analysis was required and that a simpler analysis, such as using Pearson correlation coefficients, would be inappropriate and misleading. Previously, Chambers had identified himself as one of the four reviewers of the manuscript. Many of the foregoing issues were discussed by us in the response to the original review. We also discussed these issues in a subsequent email communication with Chambers after acceptance of the manuscript. We offered to try to resolve his concerns by performing additional calculations if his concerns were stated precisely and if we agreed that suggested analyses were appropriate. No specific requests were received as of February 5, 2007.

Chambers’s last statement, “A high correlation does not automatically equate with high validity,” again used the term “correlation” which, as we discussed earlier in our response, is imprecise and likely inappropriate. It is increased agreement that supports our conclusion of increased validity resulting from application of correction for guessing.

The results and conclusions from our unique study indicate that further investigations into the utility of correction for guessing are warranted. Furthermore, we firmly believe that if instructors choose to use multiple-choice format examinations, the implementation of correction for guessing would provide better evaluation of dental students and students in other health professions as well. If these students were aware that a correction for guessing was in place, they likely would prepare more thoroughly for examinations and, as a result, their performances would improve, particularly at the lower levels of the class. But even more important than the preceding, we strongly believe that correction for guessing would result in an invaluable learning experience for students who would quickly appreciate that they need to learn not only what they know but what they do not know. We believe that students in the health professions must have a clear understanding of the limits of their knowledge and, subsequently, be able to use that understanding in their practices to the benefit of their patients.

—Thomas J. Prihoda, Ph.D.
Associate Professor of Pathology
prihodat@uthscsa.edu
R. Neal Pinckard, Ph.D.
Professor of Pathology
pinckard@uthscsa.edu
C. Alex McMahan, Ph.D.
Professor of Pathology
mcmahan@uthscsa.edu
Anne C. Jones, D.D.S.
Professor of Pathology
jonesac@uthscsa.edu
Department of Pathology
University of Texas Health Science Center
at San Antonio
7703 Floyd Curl Drive
San Antonio, TX 78229-3900

REFERENCES