Dear Dr. Alvares:

I am writing to respond to the Ranney, Wood, and Gunsolley article, “Works in Progress: A Comparison of Dental School Experiences Between Passing and Failing NERB Candidates, 2001” (March 2003). This article examined relationships between selected measures from a dental education program and pass/fail decisions on selected elements of the NERB exam for one institution. A number of concerns about the analyses and conclusions emerged from my reading of the article.

Many of the conclusions are based on an assumption of a linear relationship between variables. Conceptually, correlation is a standardized value that characterizes a non-directional relationship between data sources. Specifically, this calculated relationship captures how well each variable ranks the data in the same order. The non-directional nature of the calculated relationship is critical because it limits the conclusions that can be drawn from the results of these analyses. Within the article, the authors conclude that because there is not a statistically significant relationship between the selected institutional measures and the pass/fail decisions on selected elements of the NERB exam, this suggests there may be problems with the licensure exam. This attempt to extrapolate causality from correlation is a common mistake in data analysis. There are a number of potential explanations for these results.

First, the licensure examination may or may not be designed to measure candidates’ performance along the full ability continuum, given the purpose of the exam to focus on minimum competency with respect to entry-level skills that are necessary to protect the public. Because it is a criterion-referenced decision, rank-ordered data that provide relative position rather than attainment of specific knowledge, skills, and abilities may not be meaningful. The nature of the pass/fail decision data may also contribute to a restriction of range problem that would limit any correlation.

Second, the hypothesized relationships between these variables may or may not be meaningful because each represents a different measurement system with a different purpose. The purpose of the licensure exam was described above and represents a focused area of the content domain whereas the purpose of the other variables may serve programmatic, curricular, or instructional purposes. Because one of these measurement systems was not designed to rank order candidates (e.g., a licensure exam) and the other presumably was designed for this purpose (although the number of procedures could also be interpreted in both directions where more procedures may actually be predictive of lower ability, thus the need for more practice and vice versa), what meaning is derived from the authors’ hypothesized relationship between the two systems? A better question may be to examine the decision consistency between the two systems. This likely addresses the greater concern of accuracy.

Third, the measures from the dental education programs may or may not have validity and reliability evidence to support the use of the scores for their intended purpose(s). Because course grades or decisions may include “construct irrelevant” factors that are unrelated to the criterion-referenced definition of performance, the grades and subsequent class ranking may not be meaningful. If measurement systems are not designed to measure the same thing, any observed relationship or lack of an observed relationship may be spurious. For example, we could calculate a moderate correlation between ice cream sales and instances of violent crime in U.S. cities, but what does the relationship mean? If we infer a causal relationship, it suggests that we should ban ice cream if we want to reduce violent crime. This is a spurious relationship that demonstrates that when data are used inappropriately, the results and any conclusions drawn from them may be misleading.

The credibility of the authors’ comments about the psychometric quality of testing programs (internal or external) would be enhanced by referring to current research concerning the concepts of validity and reliability. The second edition of Educational Measurement cited in the article was replaced by the third edition in 1989, and a fourth edition is currently under development. Another common source for evaluating the quality of tests is the Standards for Educational and Psychological Testing (1999) jointly published by the American Educational Research

August 2003 ■ Journal of Dental Education
Association, American Psychological Association, and National Council on Measurement in Education. These are the professional guidelines for developing, implementing, and interpreting tests and testing programs. As in dentistry, substantive changes have occurred in the field of psychometrics over the past thirty years that are updated in more recent literature.

Although I am encouraged by the validation research that is a critical component of any measurement system, many of the conclusions drawn in this article go beyond what are reasonably supported by the reported data and may be misleading to the dental education community.

Sincerely,
Chad W. Buckendahl, Ph.D.
Director, Buros Institute for Assessment
Consultation and Outreach
University of Nebraska-Lincoln
cbuck1@unl.edu

The authors’ response:

We are happy to have the opportunity to respond to Dr. Buckendahl’s expressed concerns about our article. First, we note that he did not criticize the analytical tools we used or the results that were produced. Rather, his concerns were restricted to the conclusions we drew from them. We expect all readers to draw their own conclusions from our results, but of course were obligated to exhibit ours.

In general we find that Dr. Buckendahl has raised broad, theoretical objections to the process of relating one set of measurements to another, but he did not provide specifics on how they applied to our report. We believe, in fact, that their application validates our arguments.

Dr. Buckendahl appropriately points out that when two variables are related, it can be difficult to determine whether the relationship or lack thereof is spurious or real. He is especially concerned about our conclusions regarding the lack of a relationship between five measures of dental school performance and the restorative section (RESTOR) of the NERB exam. His first concern is that the licensure exam is a “focus on minimum competency with respect to entry-level skills that are necessary to protect the public.” We believe that his point is that dental school performance is measured on a broad set of skills, knowledge, and ability and a licensure exam has a much narrower focus. However, in addition to class rank, we used a set of measurements that were from the same content domain and should be related to the restorative section of the NERB and the simulation exercise. The RESTOR section of the exam is composed of an amalgam and composite restoration. Therefore, we tried to relate the results on that section to the number of these types of restorations done in dental school. The simulation section of the NERB relates to the preparation of simulated teeth to receive fixed prosthodontic restoration and the construction of a provisional fixed partial denture. Therefore, the measure we used for academic performance in school was the number of fixed prosthodontic units completed. The measurements we used in addition to class rank were those, given the adage “practice makes perfect,” that were likely to have a relationship with the sections of the NERB that were under our investigation.

We’re not sure how far Dr. Buckendahl might like to take an argument that licensing examinations measure something different than does the faculty of a dental school in its determination of competency for graduation. If they truly measure different things, that would put a very heavy burden on the examining agency to prove that what it measures is more important for its stated purpose (protection of the public) than what the school measures, given a lack of relationship between the two.

It would have been informative if Dr. Buckendahl had discussed his concerns in relation to our findings of a complete lack of relationship in one case (RESTOR) and a positive relationship in another case (SIM). None of our academic measures were related to the RESTOR section, yet all of them
were related to the simulation section of the test. Doesn’t this relationship in one case suggest that we have used an appropriate set of variables? At the very least, we don’t see how they could be appropriate for comparison to one section of NERB but not the other. Doesn’t the lack of a relationship between RESTOR and all of the academic variables, given a partial verification of the validity of our variables by their relationship to other variables in the NERB scores, at a minimum suggest that there could be a problem with the RESTOR section of the test?

We were motivated to conduct our investigation by increasing concerns over a number of years about the reliability of NERB’s examination. Data supporting such concern included those mentioned in the article’s discussion section for example—namely that all who persisted in taking the exam passed it within the year. This was true over all the years that NERB reported results in that format. Candidates who flunk the NERB examination at our school have to wait for the next examination, and since they are no longer students nor have a license, they have no way to increase their skills, ability, or knowledge for the retake of RESTOR. Test-retest consistency should be a feature of a highly reliable examination. On top of these concerns, we were noting large variability in pass rates year to year, much larger than easily explainable by any observed variation in academic performance among classes. Those growing concerns were compounded during the May 2001 NERB examination, during and after which the deans’ offices received numerous complaints from a number of sources about the conduct of the examination as a whole and the RESTOR exam in particular. The dean’s office asked NERB to investigate, but received a delayed, very unsatisfying response. Thus, we had a rationale, both in data and in anecdote, to suspect that there were problems with the RESTOR exam, and with our findings we believe that we have sufficient information to suggest a strengthened concern about the test.

Dr. Buckendahl’s third point of concern is especially intriguing, that is, the possibility that “construct irrelevant” factors in academic determinations may be problematic for their use in relating to reliability and validity of NERB’s test. He goes on to provide an example that is irrelevant to the material at hand. However, turning the concern around provides a very relevant example of “construct irrelevant” factors. A licensure test can also have “construct irrelevant” factors, that is, factors that are not relevant to the competency of the candidate. If, for example, the candidate’s patient fails to show for the test, or if the candidate cannot find a qualifying lesion for RESTOR due to a lack of patients exhibiting them, or if the manikin head falls off the supporting rod in SIM and breaks, then the candidate does not pass the examination and has no real recourse until the next time the examination is given, months into the future. These are clear examples of “construct irrelevant” factors, even more clearly identifiable than the theoretical ones in academic measures. One should look for additional “construct irrelevant factors” in the clinical licensing examination processes as possible causes for demonstrated lack of relationship with dental school performance and should include them in attempts to validate the tests.

Finally, Dr. Buckendahl points out that we could have used more recent references in support of discussions of validity and reliability. We agree that we could. However, he does not describe any substantive problem resulting from our use of those we cited or anything in that regard that would weaken our conclusions. We believe they stand up well.

Richard R. Ranney
Morton Wood
John C. Gunsolley