Letter to the Editor

Dental Licensure Exams: Human Subjects and One-Shot Evaluations

Dear Dr. Alvares:

I found myself and one of my colleagues quoted extensively in “A Survey of Deans and ADEA Activities on Dental Licensure Issues,” an Association Report by Drs. Ranney, Haden, Weaver, and Valachovic in the October 2003 issue of the Journal of Dental Education. The facts of our research (in Chambers and Loos, “Analyzing the Sources of Unreliability in Fixed Prosthodontics Mock Board Examinations”) were correctly reported. A single clinical test involving patients is inherently unreliable (r-values in the .10 to .20 range) and cannot be corrected by any feasible combination of improved examiner calibration or increases in the number of evaluators.

Drs. Ranney, Haden, Weaver, and Valachovic conclude that our research “would be consistent with findings of a better relationship of clinical test results in dentistry to other measures of student performance when patients are not part of the clinical test.” They interpret our paper as supporting the substitution of simulations for patient-based initial licensure testing. We did not investigate simulations, and I at least do not favor them for licensure testing.

Data from CRDTS found a lower reliability in typodont testing than with testing on live patients. The Association Report cites two studies in this matter. In one, correlation coefficients of .363 and .077 were found between class rank and pass rate in a non-patient-based licensure examination. The other study does not report the degree of association between class rank and performance on a patient-based examination or under non-patient-based testing conditions. No test was performed showing the significance of differences between these test conditions.

I would ask that the Chambers and Loos research be considered in a different light—one that urges retention of testing on live patients as part of the licensure decision. Confusion between reliability (consistency of decisions) and validity (correctness of decisions) sometimes clouds discussions of licensure testing. It is my opinion that valid determination of competency to begin independent dental practice can only be achieved based on testing involving patients because that is what dentists do. In fact such testing should be in realistic contexts, including diagnosis and management of long-term, comprehensive care.

The Chambers and Loos study was written to demonstrate that one-shot evaluation is too unreliable to support high-stakes professional decisions. The test format is correct. It just has to be used several times instead of once. The impression that rater calibration matters has been perpetuated by failure to measure the trial-to-trial variation in testing. Context-specific, patient-related variance overwhelms other considerations.

Chambers and Loos report methods for correcting this shortcoming. A valid evaluation system with low reliability can be augmented to achieve acceptable reliability through repeated testing. A testing scheme with weak validity cannot be made more valid, either through repeated testing or enhanced reliability.

The flaw in the current approach to initial licensure examination is not the patient; it is the one-shot nature of the data collection. A defensible decision about licensing candidates requires multiple measures of their performance performing comprehensive patient care.

The Association Report in the October journal is a reasoned and constructive advance over the politically defensive and psychometrically opaque debates of the past half-century of calls for initial licensure report. Any steps on that path would be progress.

I would only ask for careful consideration and clarification of a single phrase in the first of the report’s recommendations: “The criteria agreed upon unanimously by all members of the committee [AADE-ADEA Innovative Testing and Educational Methodologies Committee] are that the clinical licensure process should: 1) Be a process administered by independent third parties occurring within the educational process.” The phrase “administered by” ad-
mits of at least the two following interpretations: “conducted by” and “sanctioned by.” Expecting that third parties will come into educational programs to conduct one-shot evaluations on the same terms they do now moves the problem without solving it. If, on the other hand, boards administer the collection of reliable and valid data by setting realistic standards that schools must follow in submitting information for boards’ consideration, it will be possible to achieve confident decisions that are realistic, while also lowering cost and addressing ethics issues involving patients.

—David W. Chambers
Associate Dean for Academic Affairs and Scholarship
University of the Pacific School of Dentistry
2155 Webster Street, San Francisco, CA 94115
415-929-6438 phone; 415-929-6435 fax
dchambers@uop.edu

The authors respond:
We appreciate the attention of Dr. Chambers to our Association Report in the October issue of the Journal of Dental Education and are pleased that he found it to be a “reasoned and constructive advance” over “the debates of the past half century” about reports on initial licensure. Our sincere hope is that others will view it that way also, especially those in position to influence and initiate changes in current examination processes for initial licensure of dentists. That is why it was written.

In addition to that generally positive view of our report, Dr. Chambers described two specific concerns that he had: 1) our interpretation of the results of a cited paper of which he was senior author, and 2) wording of the first recommendation made by the AADE/ADEA ITEM Committee about the characteristics of an ideal clinical examination for initial licensure.

Prefacing the first concern, Dr. Chambers agreed that we correctly stated the facts that he and Loos reported in 1997. He disagreed with our interpretation that those facts “would be consistent with a better relationship of clinical test results in dentistry to other measurements of student performance when patients are not part of the clinical test” (quote is from our report, p. 1156). We completely understand that Chambers and Loos did not investigate simulation, and we understand that Dr. Chambers does not personally favor simulations for licensure testing. But he and Loos did conclude (p. 352) that the large variation they observed was probably due to patients.

We made our interpretation in the context of a “one-shot evaluation,” as that is how clinical licensure examinations using human subjects are currently conducted. We interpreted the results the same way that Dr. Chambers stated very clearly in his letter: namely, that using patients in a one-shot evaluation is too unreliable for high stakes tests like clinical licensure examinations. We very strongly agree with his statement that “context-specific, patient-related variance overpowers other considerations.” Those points, in fact, are among the major points we intended to reinforce by our discussion. The research by Chambers and Loos is very valuable in that regard, which is why we cited it and made it the focus of one of the ten paragraphs in our discussion section. Perhaps we should have explicitly stated that we were discussing “single evaluation” tests in order to make that point more patent.

We also agree with the contention by Dr. Chambers that validity of examinations that involve patients would be improved by multiple observations as compared to single observations. Again, that thought was inherent to our statement in the same paragraph about American Dental Association, American Student Dental Association, and ADEA

REFERENCES
policies reflecting “dissatisfaction with the current way human subjects are used.” However, that was not our entire thought, because we do differ with Dr. Chambers where he said in his letter, “The test format is correct. It just has to be used several times instead of once.” As reflected in the data from the survey of deans, there are a number of troublesome issues, including ethical ones, with respect to use of patients in the current test format (e.g., unnecessary treatment of incipient caries, treatment outside the context of a treatment plan, critical decisions made within measurement errors, etc.). We would therefore not agree that simple replication of the current test format is the appropriate solution to current problems.

We also appear to have some disagreement with Dr. Chambers on the potential of simulation in examination strategies. His letter seems to say that simulations cannot have sufficient validity and therefore should not be used at all in tests. Our point was that data show that simulations in examinations relate better to past performance in school than do determinations from single observations of restorative dentistry procedures in patients. Given that—and that, in our view, some simulations are valid for testing certain things (such as psychomotor skills)—they may well have a useful role as a part of testing for licensure.

While Dr. Chambers is correct that the cited reference number 5 did not report on the degree of association between class rank and performance on a non-patient-based examination, it did include a scatter diagram that showed no relationship between class rank and passing or failing the patient-based restorative section of the NERB examination, as well as statistical analysis revealing no difference in class rank between those passing and those failing NERB’s patient-based restorative section. The paper cited by Dr. Chambers where he said in his letter, “The test format is correct. It just has to be used several times instead of once.” As reflected in the data from the survey of deans, there are a number of troublesome issues, including ethical ones, with respect to use of patients in the current test format (e.g., unnecessary treatment of incipient caries, treatment outside the context of a treatment plan, critical decisions made within measurement errors, etc.). We would therefore not agree that simple replication of the current test format is the appropriate solution to current problems.

While Dr. Chambers is correct that the cited reference number 5 did not report on the degree of association between class rank and performance on a non-patient-based examination, it did include a scatter diagram that showed no relationship between class rank and passing or failing the patient-based restorative section of the NERB examination, as well as statistical analysis revealing no difference in class rank or GPA between those who failed and those who passed the patient-based test. In addition, a recent study¹ not yet available to Dr. Chambers found significantly better (p<0.05) mean class rank for those who passed than for those who failed the simulation sections of the NERB examination over a nine-year period, whereas there was no difference in mean class rank between those passing and those failing NERB’s patient-based restorative section. The paper cited by Dr. Chambers for information from CRDTS actually contained no data, so its statement about standard deviations and therefore putative reliability is difficult to evaluate. The other reference (Yaple et al.) indeed showed small, positive correlation coefficients between class rank and scores on the simulated clinical examination given in Ohio in 1990 and 1991, but more importantly, in contrast to the patient-based examinations given in prior years, there were fewer failures in the top quintiles of the class and more in the bottom quintiles. The authors concluded, “Compared to results from examinations given in Ohio in past years, there are fewer of the better candidates as ranked by academic standing in the dental school class who are unsuccessful on this examination.”

While we think there is evidence that well-conceived and well-designed simulations can properly be a part of clinical testing for licensure, we did not intend to say that use of human subjects under any conditions would be inappropriate. And that leads us to the second concern raised by Dr. Chambers. He took issue with ITEM’s recommendation that clinical testing should “be a process administered by independent third parties occurring within the educational process.” He explained the concern by saying that expecting third parties to come into educational programs and conduct one-shot evaluations on the same terms as they do now moves the problem without solving it. First, we stated the recommendation exactly as we did because that was the wording agreed to by unanimous vote in ITEM. Second, we completely agree with the concern as Dr. Chambers stated it and thank him for stating it the way he did because he thereby highlighted behind-the-scenes discussion on this item that occurred among the ADEA representatives to ITEM. In that discussion it was agreed that the quoted recommendation itself did not include an assumption that the process would be the same one-shot evaluations done now. The ADEA representatives to ITEM explicitly agreed to the wording as we printed it because it allows for a much different type of process than currently is used: for example, multiple observations of patient treatment under conditions that protect the patients from harm and improve both the reliability and validity of the test, as in a well-designed and well-executed portfolio approach.

We again thank Dr. Chambers for his comments, which facilitated a more complete exposition of the issues about which he wrote.

REFERENCES