Board-to-Board Consistency in Initial Dental Licensure Examinations

David W. Chambers, Ed.M., M.B.A., Ph.D.

Abstract: The consistency between student clinical performance in dental school and performance on initial licensure examinations is known to be weak. A review of the literature failed to identify any reports of the consistency between performance on initial licensure examinations and quality of technical work in practice. This research examines the consistency of performance among candidates who took two initial licensure examinations given by different testing agencies but for the same jurisdiction within a few weeks of each other. Twenty-seven candidates from one dental school took both the California Dental Board examination and the Western Regional Examining Board initial licensure examinations in 2005 and 2006. Their performance on the patient-based amalgam and composite restorations and the root planing tests were compared in these two board settings and with various dental school measures of competence. Consistent with previous findings, school-to-board performance was barely above chance levels. Board-to-board association was also insignificant and accounted for 12 percent of the common variance in the best case. Patient-based initial licensure examinations have yet to demonstrate validity in terms of consistency of performance for candidates from one performance to the next.

Dr. Chambers is Professor of Dental Education, Arthur A. Dugoni School of Dentistry, University of the Pacific. Direct correspondence and requests for reprints to him at the Arthur A. Dugoni School of Dentistry, University of the Pacific, 2155 Webster Street, San Francisco, CA 94115; 415-929-6438; dchambers@pacific.edu.

Keywords: clinical competence, clinical education, clinical evaluation, dental education, dental licensure, licensure, licensure examination, reliability, validity

Submitted for publication 1/6/11; accepted 3/7/11

It is important to get high-stakes testing right. A four-month delay caused by an unwarranted failure on the first try at one-shot initial dental licensure costs one-third year of practice income and perhaps a missed residency, private practice opportunity, or position in the uniformed services. Currently, a one-third year of practice income is equivalent to about one year of dental school debt, plus interest.1 A misclassification in the other direction entitles an incompetent new dentist to practice for the rest of his or her life, as we currently have no system for recertification.

The written National Board Dental Examination (NBDE) Part II has high internal reliability: generally above r=.90.2 The reliability of one-shot initial licensure examinations is not really known. Chambers3 has estimated the consistency from one test to another as being in the r=.40 range. This calculation was based on American Dental Association (ADA) data showing that about 97 percent of dental school graduates have an active license within one year of graduation from dental school. The r=.40 estimate is probably generous because graduates may not have an active license immediately because they are in residencies and have not taken a test, have left the country, or have not yet completed the NBDE. Verbal and informal reports from initial licensure testing organizations often place the internal consistency of their tests in the r=.70 range or even higher. This is a measure of examiner-to-examiner consistency: for any given single performance, will examiners agree with each other on the score? Examiner agreement is not a measure of consistency in candidate performance. It does not measure the consistency of candidates on multiple occasions or the consistency between performance on a licensure examination and performance in practice. Further, candidates who fail initial licensure examinations are not allowed to practice, so there is no way to determine whether the decision of the initial licensing board was accurate in those cases. Board-to-practice consistency in performance on basic technical skills is essentially unknown.

The best research available contains information about school-to-board consistency. Ranney et al.4 reported on a nine-year study of graduates of the University of Maryland Dental School, showing that there were no significant associations between class rank and performance on various sections of North East Regional Board of Dental Examiners (NERB) examination. Hangorsky5 reported that the correlation between students’ operative dentistry
clined scores in an unidentified dental school and performance on the NERB restorative section was r=.45. Correlation between students’ performance in prosthodontics courses and the restorative section of the NERB exam was r=-.009 and between dental school GPA and restorative performance on the NERB exam was r=.096. Performance on the prosthodontics section of the NERB exam was also weakly associated with performance in the discipline of removable prosthodontics in dental school (r=.110) and GPA (r=.108). Casada et al. found a correlation of r=.149 between overall performance on the Texas initial licensure examination and GPA in a sample of 372 students at the University of Texas Health Science Center at Houston Dental Branch. A four-year study of graduates of Columbia University’s School of Dental and Oral Surgery (now College of Dental Medicine) found insignificant and even negative school-to-board relationships with the NERB exam: r=-.05 for restorative, r=.11 for prosthodontics, and r=-.02 for periodontics. Although members of the dental education community tend to interpret the low school-to-board consistency as evidence that the practical licensure examination format lacks psychometric credentials, such a conclusion is not necessarily warranted. When two measurement systems designed to quantify the same characteristic disagree with each other, we can only conclude that one or both of the testing situations is inaccurate; but we do not know which is causing the problem. As medical statisticians Bland and Altman note: “When the old method has poor repeatability, even a new method which was perfect would not agree with it” (p. 149).

The research reported in this article seeks to better estimate the consistency among initial licensure examinations by studying a sample of candidates who took two different board examinations within weeks of each other. Testing purporting to measure the same patient-based technical procedures were compared to estimate board-to-board consistency.

**Materials and Methods**

A naturally occurring experiment of the type required appeared in 2005 and 2006 in California. During the early part of that decade, a series of meetings between the California State Dental Board and the dental school deans was held, the latter being critical of the format involved in one-shot initial licensure. In 2003, candidates taking the California state boards had experienced a sharp spike in failure probability. For the decade from 1997 through 2006 (except for 2003), 6 percent of candidates from the University of the Pacific, for example, experienced failure on the first testing on the California boards. This failure rate was among the lowest of the five California dental schools. In 2003, however, the rate was 14 percent. The likelihood of this being a chance occurrence was p<0.005. Similar or larger jumps in failure rates were recorded by all California dental schools that year, despite no changes in applicant pool, curriculum, or examination format. This change demonstrates that one factor affecting a candidate’s probability of passing the initial licensure examination is year-to-year standards adopted by testing agencies.

The California dental schools and the California Dental Association initiated successful legislative action that granted a California license to dentists who had passed either the California or the Western Regional Examining Board (WREB) examination, subject to passing a written test regarding the California Dental Practice Act. This resulted in a situation in which a number of students graduating in 2005 and 2006 took both the California and the WREB exams. Since 2007, the California licensure exam has not been offered to graduating dental students because the number of candidates requesting it does not justify making it available.

Three hundred ten students graduated from the University of the Pacific Arthur A. Dugoni School of Dentistry in 2005 and 2006. One hundred ninety-three took the WREB exam, with a pass rate of 96 percent; 125 took the California state board exam, with a pass rate of 90 percent. Eighteen graduates took neither examination, primarily because they moved to other parts of the country. Twenty-seven students took both the WREB and the California board exams. Scores of board candidates are made available to the school based on release documents signed by candidates. In all respects, the use and management of this dataset conform to standards for exempt clinical research protocol of the university’s Institutional Review Board.

A patient-based performance examination was included on both boards for the techniques of amalgam restoration, composite restoration, and root planing. A laboratory performance exercise in endodontics was included on the WREB exam but not the California exam. Additional tests for fixed and removable prosthodontics and endodontics were
part of the California exam, but are not considered here because they were not patient-based. A written periodontics diagnosis and a prosthodontics written test were included on the WREB exam and are not considered in this article.

In-school, patient-care performance indicators were obtained from students’ academic records. Scores on the board and school measures were matched using a secure code number. The school performance measures included grade point average (GPA) in lab and clinical courses and average scores on test cases, average faculty ratings, and total number of procedures performed in each of the disciplines for which licensure exam scores were available on patient-based tasks.

The primary statistic used in this study was the Pearson correlation coefficient, comparing, for example, the score students received for the amalgam test on the WREB and California examinations. Because correlation coefficients are not expressed in linear units (an r-value of .40 is not twice as large as an r-value of .20), the discussion of results is framed in terms of the coefficient of determination. This is simply the r-value squared, conventionally represented R², and is the proportion of variance shared by the two measures; it is the proportion of all information one has to what would be needed to predict with 100 percent confidence the value of one variable given the known value of another. For example, the length of the mandible measure in inches provides 100 percent of the information needed to know the length of the same mandible measured in centimeters: r=1.0, R²=1.00, proportion of known variance=100 percent. The correlation coefficient reflecting consistency among licensure examiners looking at a single candidate performance is said to be about r=.70, R²=.50. Knowing one examiner’s score provides about half of what one would need to know to accurately predict the other examiner’s score. However, the portion of what is needed to predict the candidate’s performance on the next trial of that procedure cannot be determined from this value.

A power test was performed setting r=.70, α=.05, and 1-β=.80. A sample size of twenty would be sufficient for a test of the hypothesis that the correlation between test scores on alternative initial licensure examinations is at least as high as r=.70. The sample size of thirty-seven from the naturally occurring experiment reported here is thus almost twice the required size.

Results

Table 1 shows the measures of association (correlation coefficients) for the WREB and California licensure examination performance of students from two graduating classes on three patient-based sections, one laboratory exercise, and various predictive values. The first column (board-to-board) is the

| Table 1. Correlation coefficients for patient-based performance on sections of initial licensure examinations and various predictive factors |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Amalgam                        | WREB            | .325            | -.014           | .016            | .064            | .017            |
| California                     | .294            | -.022           | .064            | .260            |                |
| Composite                      | WREB            | .341            | .062            | .052            | .063            | -.074           |
| California                     | .152            | .226            | .045            | .087            |                |
| Root planing                   | WREB            | -.029           | .124            | -.018           | .062            | .069            |
| California                     | .056            | -.144           | -.084           | .017            |                |
| Endodontics                    | WREB            | .292            | .181            | .147            | .059            |
|                               |                 |                 |                 |                 |                 |

Brd-to-Brd=correlation between same sections of initial licensure examinations offered by two testing agencies to same set of students (n=27).
Clin GPA=cumulative grade point average of candidates while students are in dental school.
Test Case=average score on test case simulations of examined patient-based disciplines.
Fac Rating=final quarter ratings by faculty members of students’ competence in various tested disciplines.
N Proc=number of procedures in relevant disciplines completed by candidates while in dental school.

Note: None of these associations is statistically significant at p<0.05, two-tailed test.
consistency between scores for students who took both the WREB and California examinations. The next four columns display associations between board performance and school-based performance in these disciplines. The school-based predictors include GPA in all lab and clinical courses, average test case scores (independent board simulations) in the same areas as those of the licensure examination, average of faculty ratings in the final quarter for the discipline tested on the board, and number of procedures completed by students during their educational program in each discipline.

None of the correlations shown in Table 1 is statistically significant at $p<0.05$ with a two-tailed test. There is no evidence of consistency between boards on clinical tests of the same procedures, and there are no in-school predictors of performance on initial licensure examinations.

**Discussion**

The results of this investigation are consistent with findings previously reported in the literature. None of the school-to-board correlations in the present study were significant, and the largest (an $r$-value of .294 between overall clinical GPA and score on the California examination in amalgam restorations) explains less than 10 percent of the variance in board performance. In addition to the previously reported weak predictive value of clinical GPA and faculty ratings, average scores on test cases that simulate independent performance in situations that mimic licensure exams and amount of experience (number of cases completed) also fail to provide predictive information. None of these predictors is consistently better than the others, and all leave substantial amounts of unpredictable, random variation in the licensure testing context.

The more important contribution made by this study is to provide an estimate of the consistency that can be expected from one initial licensure testing situation to another (board-to-board consistency). The testing situations were a matter of weeks apart and may thus be regarded as independent tests of the same patient-based ability to begin independent dental practice. Chambers reported on a method for estimating this value based on national data regarding overall failure rates and placed the equivalent of the correlation coefficient at $r=.400$. In the current sample, the observed correlation coefficients were $r=.325$ for amalgam restoration, $r=.341$ for composite restoration, and $r=-.029$ for root planing.

In this study, the estimate of consistency within the testing situation is at best less than 12 percent of the common variance. This means that more than 85 percent of the factors that account for student performance differ across the two initial licensure situations. Because the candidate's ability to practice is, per definition, understood to be common across such temporally proximate and procedurally similar examination settings, it is evident that extraneous factors play a large role in one-shot, initial licensure examinations. Examiner consistency is unlikely to account for a large segment of this random variation, a conclusion based on the care that testing agencies take in calibration. This is a discouraging conclusion since it is unlikely that further efforts at improving calibration will matter much. The most reasonable explanations for unaccounted variance include patient and situational variability. The presence of such sources of variance in dental education has already been reported. It is also plausible that initial licensure examinations, in an effort to boost accuracy, are actually oversensitive to minor variations that would not be material in distinguishing the work of one practicing dentist from another. Only the variation at the extreme low end of performance should matter in the initial licensure decision.

Some potential limitations in this research are worthy of discussion. First, there is the reflexive concern over sample size. The conventional power analysis shows that this study is more than adequately powered. Sample size is, however, related to the plausibility of alternative correlation coefficient values. The range of such reasonable estimates is conventionally expressed as a confidence interval. The best guess based on the data is that the observed coefficient of determination for consistency between WREB and California initial licensure examinations on composite restorations is 12 percent. But unlike $t$-tests, the confidence intervals for coefficients of determination are not symmetrical around the "best point estimate." The probability that the coefficient of determination is actually as high as 50 percent involves calculating the standard error of Fisher $z$-transformations, a correction that is necessary because correlation coefficients have skewed distributions. The chance that the true value for the coefficient of determination is .50 as claimed is $p<0.0039$. The same procedure is followed in calculating the probability that the relationship between performance on the different boards
is an absolute zero. In this case, the probability that there is no association between performance on the two boards tested in this study is $p<0.058$. The claim that there is no association between performance on the same clinical procedures on two boards is fifteen times more likely to be true than the claim that at least half of the information needed to predict future performance can be determined from a single performance on an initial licensure examination.

There is also the possibility that the WREB and California testings were not independent of each other. Perhaps students signed up for both and only took the second if they failed initially. However, this research found that 15 percent of the candidates who took both tests failed the WREB exam (the examination given at the earlier date) and subsequently took and passed the California exam. The proportion of candidates who passed the WREB exam and subsequently took the California exam is a very similar 12 percent. The hypothesis of selective second examination therefore does not appear plausible based on the data.

Research by Chambers and Loos suggests that in-school consistency across multiple simulations of licensure testing will be low just as it is for board-to-board testing. It appears that patient-based assessment of technical procedural characteristics is inherently prone to high variability across testing instances, whether performed in school settings or elsewhere. The problem is that a one-shot approach is too small a work sample to justify a confident assessment of student or candidate ability. It is a law of measurement theory that consistency increases automatically when repeated independent measures are taken, even when the consistency of each measure individually is not especially strong. Only evaluation decisions made on the basis of multiple performance samples in realistic settings have the potential to provide estimates of ability needed for high-stakes decisions.

Conclusion

This research is the first to address the question of board-to-board consistency, the likelihood that candidates who perform well on one administration of an initial dental licensure examination will perform well the next time they execute this procedure in a similar setting, such as another board examination or in practice. It was found that board-to-board consistency on patient-based amalgam and composite procedures accounts for about 12 percent of the variance within the same candidate but across trials. This estimate is similar to but slightly smaller than the 16 percent estimate of overall board-to-board pass rate in a previous study, but much better than the 1 to 2 percent estimates in the literature for school-to-board consistency. It appears that there was no consistency between one board testing and the other on the patient-based root planing procedures.

The finding of low board-to-board consistency is relevant to claims that might be made about board-to-practice consistency. Bland and Altman, in the argument presented at the beginning of this article, state that if either part of a claimed association is known to be weak, it cannot achieve high predictive validity for the other measure. There are certainly differences between the content and format of the California and WREB examinations, but both claim to measure exactly the same thing: the candidate’s suitability for independent practice in a particular state.

If initial licensure testing agencies wish to demonstrate their validity as predictors of performance following the one-shot testing format currently used, additional psychometric studies will be needed that show patterns of performance-to-performance consistency that more closely approximate the 50 to 75 percent common variance now accepted as the standard among high-stakes tests. Alternatively, decisions could be made based on multiple assessments in realistic settings, where a diverse array of context-embedded evidence of performance replaces the one-shot testing, regardless of whether done in dental schools or elsewhere.

REFERENCES