The relationship between performance in a dental school and performance on a clinical examination for licensure: A nine-year study
RICHARD R. RANNEY, JOHN C. GUNSOLLEY, LOIS S. MILLER and MORTON WOOD
J Am Dent Assoc 2004;135;1146-1153

The following resources related to this article are available online at jada.ada.org (this information is current as of February 27, 2009):

Updated information and services including high-resolution figures, can be found in the online version of this article at:
http://jada.ada.org/cgi/content/full/135/8/1146

Information about obtaining reprints of this article or about permission to reproduce this article in whole or in part can be found at:
http://www.ada.org/prof/resources/pubs/jada/permissions.asp
The relationship between performance in a dental school and performance on a clinical examination for licensure: A nine-year study


Clinical testing for licensure has come under increasing scrutiny in recent years. Concerns about the process include validity of the examinations for licensure decisions, ethical and other issues in the use of live patients, and large variation in failure rates among examinations given by different testing agencies. One might expect a positive relationship between performance while a student is in a dental educational program and performance on clinical licensure examinations. However, published data do not uniformly support that conclusion.

A recent report found no differences in class rank or grade point average, or GPA, between graduates who failed and those who passed the restorative section (amalgam and composite restorations) of a clinical examination given by the North East Regional Board of Dental Examiners, or NERB. This report also found a wide distribution of class ranks for both those who failed and those who passed NERB. At the same time, there was a difference in academic performance between students between those who passed and those who failed a NERB exercise on a manikin, though again the distribution of

Background. Licensure examinations in dentistry have become an increasing concern, owing to ethical issues in the use of patients, difficulties in seeing relationships between outcomes of licensure examinations and performance in educational programs, and questions on the reliability of “one-shot” clinical examinations. Using data from a nine-year period, the authors compared the results of clinical licensing tests and the academic class ranks of the candidates.

Methods. The authors studied data for 835 dental school graduates of one school from 1994 through 2002. They compared the dental graduates’ results from the North East Regional Board, or NERB, of Dental Examiners examinations with their class ranks. The authors used analysis of variance to analyze the differences among passing, failing and “no data” groups, K statistic and logistic regression for variation, and receiver operating characteristic, or ROC, curves for diagnostic utility.

Results. The class rank of graduates who passed and failed NERB’s restorative section of the examination did not differ. Differences for other sections of the examination were statistically significant but small. The variation in restorative and manikin exercises over time was highly significant. No consistency existed between these tests, and their ROC curves indicated no utility for diagnosing class rank.

Conclusions. The authors’ analysis of nine years’ data called into question the reliability and validity of initial licensure examinations based on certain of the one-time tests used by NERB. Future study should determine if the results generalize to other schools and clinical testing agencies.

Practice Implications. If the results of this study can be generalized to all U.S. licensure examinations, basing licensing decisions on clinical licensure examination alone risks licensure decisions of low validity. Use of patients in examinations of questionable validity may be unethical because they may have been subjected to risk of irreversible damage without contribution to a valid decision-making process by the licensing authority.
class ranks among failing and passing graduates was highly dispersed. The results of that study added to questions about validity of licensing examinations for making decisions about licensure and increased concerns about the ethics of irreversible procedures on patients in those tests. Those results, however, were for a single year only, which might have been unrepresentative of the general results of NERB’s examination. Therefore, in the current study, we assessed the relationship between dental students’ performance in dental school and performance on NERB’s clinical examination by assessing results over nine years.

METHODS

We studied the results of NERB clinical examinations that were performed in May of the years 1994 through 2002 at the Baltimore College of Dental Surgery, Dental School, University of Maryland. We analyzed data representing the 835 doctor of dental surgery graduates of the school during that period. We determined the class rank for each graduate within each class based on his or her overall GPA, and then we normalized the class rank for comparability among classes by converting it to a percentile. We used the results of each graduate’s first time taking NERB’s clinical examination (as reported to the school by NERB) from each of the examinations major sections:

- the dental simulated clinical exercise, or DSCE (written);
- the restorative clinical exercise, or RESTOR;
- the simulated patient treatment (manikin) clinical exercise, or SIM PATIENT;
- the periodontal clinical exercise, or PERIO.

Since NERB uses a conjunctive scoring method, the overall result was failure for those who scored below 75 on any of the sections.

RESULTS

Figure 1 depicts the year-to-year variation in failure rates over the nine years of this study. The overall failure rates, the RESTOR section and the
and 10 percent passed the SIM PATIENT section but failed the RESTOR section. The K statistic for the comparison was 2 percent, which was lower than its standard error of 4 percent.

The total numbers of graduates who passed, failed or had no reported results from the NERB examination in their respective graduation years (1994-2002) appear in Table 2. Mean class rank percentiles for those who passed, failed or had no reported results are shown in Table 3. Nine-year failure rates were 4 percent for the DSCE section, 6 percent for the PERIO section, 10 percent for the RESTOR section and 13 percent for the SIM PATIENT section. Conjunctive scoring produced an overall 29 percent failure rate on initial attempts to pass NERB’s clinical examination; this was more than twice the rate of any single section. If the failure rate of the four sections were to be summed, conjunctive scoring would have meant that 33 percent of the graduates having reported results would have failed the overall test if none of them failed more than one section. Therefore, failing more than one section of the test was a rare occurrence (4 percent of all, and only 13 percent of those who failed at least one section).

We detected no statistically significant difference in class rank percentile between those who passed and those who failed the RESTOR section, though those who passed differed from those who had no reported results. In the overall passage/failure results and in those for the SIM PATIENT section, those who passed had a lower (better) class rank percentile than did either those who failed or those with no reported results. While those differences were statistically significant, they were numerically small and close to the median. The group that passed the PERIO section had a better class rank percentile than either those who failed or those with no reported results. And, the DSCE section showed the largest distinction in class rank percentile between the passing and failing
groups. This was owing to a worse class rank for the failing group for this section compared with the other sections of the examination.

The ROC curve for evaluating class rank by failure of the NERB clinical examination (overall failure) is shown as Figure 2. It indicates that the examination was not a good diagnostic tool for that purpose. The curve is close to the diagonal, and there is no point on the curve that has high sensitivity and an acceptable false-positive rate (one minus specificity). Each of the ROC curves for the sections involving on-site evaluations by examiners was similar to the curve for overall results. To illustrate, the ROC curve for the NERB’s RESTOR section is presented as Figure 3. The analogous curves for the PERIO and SIM PATIENT sections were nearly the same, so we did not include illustrations for them. Only the DSCE section offered the possibility of achieving a 90 percent sensitivity at less than a 60 percent false-positive rate (Figure 4, page 1151). At 80 percent sensitivity, the DSCE section had about 30 percent false-positives, and at 70 percent sensitivity, it had about 15 percent false-positives.

**DISCUSSION**

The most important feature of any test is the degree to which it provides a basis for a valid decision.15 As reflected in this study, the clinical test administered by NERB over a nine-year period to students from one dental school exhibited a number of characteristics that can be used to argue against its validity for decisions by licensing authority on whether to grant a license to practice dentistry.

The significant variation in certain failure rates from year to year suggests that either the tested abilities of graduates were different from year to year or that the NERB examination itself was different from year to year. Concern about this variation in the test is compounded by the inconsistency between the results of the SIM PATIENT and the RESTOR sections within the

---

**TABLE 2**

<table>
<thead>
<tr>
<th>NERB SECTION</th>
<th>NO. OF GRADUATES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Passed</td>
</tr>
<tr>
<td><strong>RESTOR</strong>†</td>
<td>536</td>
</tr>
<tr>
<td><strong>SIM PATIENT</strong>‡</td>
<td>519</td>
</tr>
<tr>
<td><strong>PERIO</strong>§</td>
<td>561</td>
</tr>
<tr>
<td><strong>DSCE</strong>¶</td>
<td>529</td>
</tr>
<tr>
<td>Overall</td>
<td>417</td>
</tr>
</tbody>
</table>

* NERB: North East Regional Board.
† RESTOR: Restorative clinical exercise.
‡ SIM PATIENT: Simulated patient treatment (manikin) clinical exercise.
§ PERIO: Periodontal clinical exercise.
¶ DSCE: Dental simulated clinical exercise (written).

**TABLE 3**

<table>
<thead>
<tr>
<th>NERB SECTION</th>
<th>MEAN CLASS RANK PERCENTILE†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Passed</td>
</tr>
<tr>
<td><strong>RESTOR</strong>†</td>
<td>46*</td>
</tr>
<tr>
<td><strong>SIM PATIENT</strong>‡</td>
<td>45*ab</td>
</tr>
<tr>
<td><strong>PERIO</strong>§</td>
<td>46*ab</td>
</tr>
<tr>
<td><strong>DSCE</strong>¶</td>
<td>45*</td>
</tr>
<tr>
<td>Overall</td>
<td>42*ab</td>
</tr>
</tbody>
</table>

* NERB: North East Regional Board.
† Percentages in the same row with the same superscript letter differ from each other, \( P < .05 \).
† RESTOR: Restorative clinical exercise.
‡ SIM PATIENT: Simulated patient treatment (manikin) clinical exercise.
§ PERIO: Periodontal clinical exercise.
¶ DSCE: Dental simulated clinical exercise (written).
NERB examination. Although there are differences in some skills tested in the SIM PATIENT and RESTOR sections (for example, the SIM PATIENT section uses a typodont for an extracoronal procedure and the RESTOR section requires clinical decision making for an intracoronal procedure in a patient), they are the most closely related of the four components of the NERB examination. The final products involve preparations and physiologically contoured restorations that use similar hand-eye coordination skills. Consequently, one would expect some measure of agreement between them. But with a K statistic that was essentially zero, the only possible conclusion is that the tests fail to validate each other. The finding that failing more than one section of the examination was a rare occurrence strengthens this conclusion. These findings support a hypothesis that the difference in failure rates over the nine years was related to inconsistent evaluations by the clinical evaluators, not to variation in the abilities of the graduates over those years. The hypothesis also is supported strongly by the lack of variation over time of the PERIO and DSCE sections of the evaluation. Whereas NERB and other clinical testing agencies do strive for intraexamination reliability by standardization exercises for examiners, the results of our study indicate that the interexamination reliability (year to year) is not good and that the examiners are not consistent among the different sections of the NERB. We are aware of no other published analysis of these types of variations in clinical dental examination results.

This is not to say that the standardization exercises are without value. Standardization should reduce variation due to measurement error. Standardization for examiners, however, does not ensure that the overall test is valid, or that other, even larger, sources of variation are controlled. In fact, variation attendant to the use of nonstandardized patients as part of the examination can be substantially larger than variation attributable to measurement error, thus reducing or destroying the reliability of the clinical test.

Decisions for licensure should be based on tests that are both valid and reliable. If the variability found in our study is representative of tests in other licensing jurisdictions, decisions across the nation about licensure are being made by licensing authorities on the basis of observations of clinical testing agencies that are
suspect for reliability and validity. There is no question that the dental examiners for these testing agencies are dedicated people who take time from their practices or other professional and personal pursuits to conduct the examinations for the betterment of the profession and protection of the public. Despite their efforts, however, the data in this report indicate that the NERB examiners are not likely to accomplish their goal of eliminating unqualified people from licensure. Over the time that NERB has reported results as those who “availed themselves of all opportunities to pass the NERB Clinical Examination in Dentistry,” 100 percent of the graduates in our study passed (E.H. Hall, director of examinations, North East Regional Board, written communications, Jan. 23, 2002, and Jan. 15, 2003). NERB’s failure to reach the same conclusion on first examination came at the cost of denying licensure to competent graduates for some period during a time when their educational debt burden is at an all-time high.

Over a nine-year period, there was no significant difference in class rank percentile between those who passed the RESTOR section of the NERB examination and those who failed it. This indicates that a one-time evaluation by NERB examiners of restoration preparation, caries removal, and placement and finish of amalgam and composite restorations essentially does not relate to the quality of the respective students as determined by the dental school faculty. This finding is in agreement with a previous report from a single year’s results from an examination given by NERB. As the faculty’s determinations are based on multiple observations, and validity of decisions is improved by use of multiple observations, the usefulness of the NERB examiners’ determination that a graduate lacks competence in restorative dentistry is questionable. On the basis of the data from our study, one can conclude that the state boards of dental examiners should question the clinical licensure examiners’ conclusion in that regard, and take more seriously the determination of the faculty. To assure the public that there is not a conflict of interest for the faculty in determining qualification for practice, perhaps those making the licensure decision should take both the faculty’s observations and the observations of an independent third party into account. But based on the results of our study, a decision should not be based solely on the determination of the examining agency, as the decision would in that case lack sufficient validity.

From the data, NERB’s use of conjunctive scoring clearly elevated the failure rates by more than double the rate of any of the examination’s contained sections. In selecting conjunctive scoring, NERB argues that a passing mark in each section is necessary to ensure protection of the public by independent evaluations of competence for each part of the examination that they determine to be important. If reliability of the examiners’ determinations was good, that assertion would be plausible. However, it is not realistic to accept that argument, as reliability from section-to-section was nonexistent when we evaluated it over a nine-year period, and major sources of variation that can contribute to a failing score remain uncontrolled. It would appear, in fact, that the conjunctive scoring method decreases the reliability of the pass/fail decision at the level of the examining agency. Therefore, it also decreases the validity of the decision at the level of the licensing authority if the licensing authority accepts the agency’s evaluation without considering other factors.
Even for the PERIO and SIM PATIENT sections of the NERB examination for which mean class rank percentiles did differ significantly between pass and fail results, those differences were not large; they were only 45 or 46 versus 58, a difference of 12 to 13 percentile ranks out of a possible 99. So while these differences were statistically significant at \( P < .05 \), it is unlikely that they were significant in terms of validity of decisions made on the basis of them.

Our conclusion that NERB’s clinical examination lacks reliability as a requirement for licensure is supported further by the ROC curves produced from the data in our study. The ROC curves demonstrate that if the intention is to detect the poorest performers in the graduating classes, the clinical tests do not do the job. The examining community has asserted that only a small percentage of graduates (perhaps 2 to 3 percent) should be prevented from obtaining a license to practice through the clinical testing process. It seems reasonable that the worst 2 to 3 percent of the graduates might be found in the lower portion of the dental school class as determined by the dental school faculty. But our ROC curves showed that NERB’s clinical tests could not do much better than a random possibility of making that determination. Most people who failed the examination on their first try did not reliably deserve to. And at least for those years for which we have the relevant report from NERB, all the graduates who persisted in taking the examination after failing the first time did pass within the same year.

Over the nine years we studied NERB’s DSCE section, it had a 33 percent rank differential between candidates who passed or failed, which was between double and triple the differential for the clinically evaluated sections. Its ROC curve also indicated that it came closer to being diagnostically useful for academic performance than any other section or the overall results of the NERB examination. We expect that a substantial part of the reason for this is because the uncontrolled variation attendant to use of human subjects (patients) in the RESTOR and PERIO sections, and the subjective determinations made in those sections and the SIM PATIENT section are not present in the DSCE section. It also is possible that part of the reason is that the DSCE section is more analogous to the type of grading and ranking most commonly encountered by students in dental school.

Most, if not all, of the jurisdictions that use the NERB examination also require a passing score on Parts I and II of the National Board Dental Examinations. In addition, some dental schools, including the school that was the source of data for our study, require that a student pass that examination before graduation. The natural question is whether passing both the National Board Dental Examination and NERB’s DSCE section is a reasonable requirement for licensure. A comparison of NERB’s DSCE section and Part II of the National Board Dental Examination performed at the request of the ADA House of Delegates in October 1998 reportedly concluded that they measure different things. That comparison could not determine, however, whether one examination provided more useful information than the other for purposes of the licensure decision, or whether either examination would identify the same people as having passed or failed. A direct comparison of students’ in-school performance on Part II of the National Board Dental Examinations with their performance on NERB’s DSCE section showed that the results from both examinations essentially were the same.

While our current study improved on previously published data by using results over a number of years, it still was limited to one school, meaning also that it was limited to that school’s educational program and its facility as a NERB examination site. It would be useful to conduct similar analyses of data from several schools together and from different examining agencies.

**CONCLUSION**

Over a nine-year period, the NERB examination results of graduates from one dental school failed to be a good measure for detecting the quality of those graduates as determined by the dental school’s faculty. The sections of the NERB examination that were dependent on examiner observations were less able to make a distinction between good and poor class rank than was NERB’s DSCE section. The interexamination reliability of the NERB examination was low, as indicated by the high year-to-year variation in the clinical examination results and the fact that different sections...
of the examinations were not able to validate each other, while the results from the DSCE section did not significantly vary from year to year. The clinical examinations did not provide validity for making the licensure decision, bringing into question the ethics of using invasive and irreversible procedures on patients as a part of the dental licensure examination.

Dr. Ranney is a professor of periodontics and a former dean, Baltimore College of Dental Surgery, Dental School, University of Maryland, and is the Gries Education Fellow, American Dental Education Association. Address reprint requests to Dr. Ranney at Baltimore College of Dental Surgery, Dental School, University of Maryland, 666 W. Baltimore St., Baltimore, Md. 21201, e-mail “rranney@dental.umaryland.edu”. Address reprint requests to Dr. Ranney.

Dr. Gunsolley is a professor and the chair, Department of Endodontics/Periodontics, Baltimore College of Dental Surgery, Dental School, University of Maryland.

Ms. Miller is the director, Academic Support Services, Baltimore College of Dental Surgery, Dental School, University of Maryland.

Dr. Wood is an associate professor and the chair, Department of Restorative Dentistry, Baltimore College of Dental Surgery, Dental School, University of Maryland.