Validating the Standard for the National Board Dental Examination Part II

Tsung-Hsun Tsai, Ph.D.; Laura M. Neumann, D.D.S., M.P.H.; John H. Littlefield, Ph.D.

Abstract: As part of the overall exam validation process, the Joint Commission on National Dental Examinations periodically reviews and validates the pass/fail standard for the National Board Dental Examination (NBDE), Parts I and II. The most recent standard-setting activities for NBDE Part II used the Objective Standard Setting method. This report describes the process used to set the pass/fail standard for the 2009 exam. The failure rate on the NBDE Part II increased from 5.3 percent in 2008 to 13.7 percent in 2009 and then decreased to 10 percent in 2010. This article describes the Objective Standard Setting method and presents the estimated probabilities of classification errors based on the beta binomial mathematical model. The results show that the probability of correct classifications of candidate performance is very high (0.97) and that probabilities of false negative and false positive errors are very small (.03 and <0.001, respectively). The low probability of classification errors supports the conclusion that the pass/fail score on the NBDE Part II is a valid guide for making decisions about candidates for dental licensure.

Dr. Tsai is Manager, R&D/Psychometrics, Department of Testing Services, American Dental Association; Dr. Neumann is former Senior Vice President, Education, American Dental Association; and Dr. Littlefield is Director, Academic Center for Excellence in Teaching, University of Texas Health Science Center at San Antonio. Direct correspondence and requests for reprints to Dr. Tsung-Hsun Tsai, Department of Testing Services, American Dental Association, 211 E. Chicago Ave., Chicago, IL 60611; 312-440-2684 phone; 312-587-4105 fax; tsait@ada.org.

Keywords: educational measurement, assessment, dental licensure, objective standard setting, NBDE Part II

As part of an overall exam validation process, it is necessary to objectively set the standard or passing/failing score. Discussion and research on setting standards for licensure and certification examinations suggest that objectivity is a critical step in the validity of high-stakes certification and licensure decisions.¹⁻⁴ This article describes the standard-setting procedure used by the Joint Commission on National Dental Examinations. Using a change in the standard for the 2009 exam, the impact on the number of false positive and false negative decisions is estimated using the beta binomial mathematical model.

The Joint Commission periodically conducts standard-setting activities to objectively establish the minimum passing score on the National Board Dental Examination (NBDE), Parts I and II. Setting the pass/fail score involves both empirical data and the value judgments of experts. The most recent standard-setting activities for the NBDE Part II were conducted in 2008. The Joint Commission noted that failure rates increased from 5.3 percent in 2008 to 13.7 percent in 2009 after the new standard was applied. This increase concerned dental educators and students. A formal review of scoring, score reporting, and standard-setting activities was therefore conducted. These processes were also reviewed by external consultants. The Joint Commission concluded that scoring, score reporting, and standard-setting processes were appropriate and followed guidelines in the Standards for Educational and Psychological Testing.⁵

The Joint Commission communicated these findings to communities of interest through a variety of channels, including presentations at professional conferences, letters to dental educators and the leadership of the American Student Dental Association, and the Joint Commission Newsletter. This article describes the Objective Standard Setting process used for the NBDE Part II and presents estimated probabilities of decision errors based on the standard using actual data from a 2007 administration of the NBDE Part II.

NBDE Part II

The NBDE Part II is designed to assist state dental boards in determining the qualifications of individuals who seek licensure to practice dentistry. The examination is typically taken by student candidates during their last year of dental school. It assesses the ability to understand important information from biomedical science, clinical dental science, behavioral science, and patient management and the ability to apply such information in a problem-solving context.
context. This comprehensive, computer-based examination consists of 500 multiple-choice items. Items include sample content from multiple disciplines intermingled throughout the examination. Items are presented with a stem pairing a question or statement with a list of four or five possible responses. The examination consists of 400 discipline-based items and 100 items based on eight to ten case scenarios. Each case presented in the examination consists of a synopsis of a patient’s health and social histories, patient dental charting, diagnostic radiographs, and clinical photographs.

Panelists for confirming the standard for the NBDE Part II were selected by following guidelines from the Standards for Educational and Psychological Testing. The Joint Commission reviewed and approved the selection criteria. Seven panelists, who were selected and appointed by the Joint Commission, had the following characteristics: they represented all major regions of the United States, they included full-time practitioners and dental educators affiliated with accredited dental education programs, and they included subject matter experts from multiple content specialties.

Standard-Setting Procedure

The standard-setting activities were conducted using the Objective Standard Setting method. On-site training was provided to panelists at the standard-setting meeting, including the panelists’ role and responsibilities, background and purpose of the NBDE Part II, a brief review of all meeting materials, and steps involved in confirming the standard. Specific training was also provided to the panelists. A standard-setting protocol was developed to provide detailed information regarding concepts and the use of the Objective Standard Setting method. During the training, the protocol was reviewed.

To help the panelists conceptualize and correctly use the Objective Standard Setting method before the actual standard-setting activities, sample items from the NBDE Part II were used. During the practice process, each panelist rated each sample item individually by judging its importance to patient care using a rating scale ranging from 1 to 5, with 1 indicating very unimportant to patient care and 5 indicating critically important to patient care. After the ratings were complete, the group was asked if there were any specific problems or issues that arose while they were rating the item. Their concerns or issues were then addressed and discussed. Based on the group discussion, panelists were given the opportunity to change their ratings if they would like to. Once the discussion and revisions were done, the group moved on to the next sample item. This process was repeated until the panelists understood the concept and felt comfortable to apply the concept to the real activities.

The Objective Standard Setting method requires panelists to make three recommendations: 1) a selection of criterion items and proportion of these items that minimally competent candidates would answer correctly; 2) a determination of the percentage of mastery; and 3) a determination of the confidence level. Each panelist selected items that he or she considered to be very important using the following criteria:

- The content of criterion items must be central, or directly related, to practice.
- Criterion items must assess the knowledge and problem-solving skills that are employed frequently in practice.
- Criterion items must assess the knowledge and problem-solving skills that are subject to change with current research and development in the field.
- The content of the criterion items must be of fundamental and critical importance to successful practice.
- Criterion items must be selected from a full range of the content included on the examination.

The next type of information needed was related to the level of mastery. The panelists were instructed to record their estimates for an acceptable level of mastery and indicate which level of mastery (0 to 100 percent) should be necessary to pass the NBDE Part II. This estimate should be based on the panelists’ knowledge of the reference group and the overall field sampled by the examination. The reference group consists of all students who are currently enrolled in accredited dental schools and who are taking the examination for the first time.

Finally, the information regarding the extent of error was necessary to complete the standard-setting activities. The panelists recorded their estimates as to how large the error band around a score should be. The notion of error is involved in understanding the performance of candidates. The true score of a candidate is somewhere within an error band. When a candidate’s score falls within the error band around
the standard, the question is whether the score should be evaluated as a passing or failing score. Several considerations are important here. First, does one wish to emphasize protecting the public? If so, one would pass only candidates whose scores exceed the upper limit of the error band. Second, if one wishes to focus on protecting the candidate, one would pass all candidates whose scores exceed the lower limit of the error band. Finally, the question arises about how large the error band should be. According to Kramer and DeMarais, a 95 percent confidence level is considered appropriate.

**Standard-Setting Results**

Based on the panelists’ judgments, the passing/failing score was set using the Objective Standard Setting method, and the score scale was established using the Rasch model. In the Rasch model, candidate ability and item difficulty are described by a measurement scale. This means that candidate ability can be directly related to the specific abilities, knowledge, and problem-solving skills that underlie items on the NBDE Part II. The candidate ability is computed based on the probability of a right or wrong response on each item. The underlying ability scale is centered on zero and typically ranges from a -5.00 to a +5.00, with more negative values indicating relatively easier items and lower-scoring candidates. In like manner, more positive values indicate relatively more difficult items and higher-scoring candidates. Because candidate ability and item difficulty are on the same scale, it is possible to directly relate the two statistics relative to the criterion items. According to the judgments of panelists, the knowledge underlying the criterion items is critically important to patient care. The passing rule sets the cut score at the average difficulty of these criterion items, and the cut score is set at that point along the scale. Those candidates whose score is at or above this point would pass. This point along the measurement scale is assigned a standard score of 75.

After the minimum passing score was determined through the standard-setting activities, the abilities of candidates were estimated for every possible raw score (number of correct responses), ranging from 0 to 500. Score conversions were developed to translate raw scores into standard scores for all exam forms using the common-item equating design. Outcomes from the 2008 standard setting activities were applied in 2009.

**Evaluation of Standard-Setting Outcomes**

To evaluate the outcomes from the standard-setting activities from 2008, the new standard from the panelists’ recommendation was also applied to the 2007 examination results. In 2007, 4,464 candidates in the reference group took the NBDE Part II. The reference group is comprised of all students enrolled in schools with approval accreditation status who took the examination for the first time. If the new standard had been employed in the 2007 examination, 7.1 percent of the candidates in the reference group would have failed instead of 6.4 percent, i.e., thirty-one additional failures. This increased failure rate of 0.7 percent is less than the increased national failure rate between 2008 (5.3 percent) and 2009 (13.7 percent).

In addition to examining the impact of the new minimum passing score on failure rates, research was also conducted on the classification accuracy of the pass/fail score. Classification accuracy is the extent to which decisions would agree with those that would be made if candidates could be tested with all possible forms of NBDE Part II (Table 1). In Table 1, correct classifications occur when the decision made based on the true score (i.e., average score of a candidate that took all forms of the examination) agrees with the decision based on the observed score (exam form actually taken). False positive and false negative classifications refer to mistakes: a mismatch between a candidate’s true score and observed score. The overall false positive value is the proportion of candidates who received an observed score of “Pass” but their true score was a “Fail.” Likewise, the overall false negative value is the proportion of candidates who received an observed score of “Fail” but their true score was a “Pass.” This approach for analyzing classification accuracy is routinely used in licensure testing.

Data used in this analysis were from the reference group of 1,252 candidates taking one form of the NBDE Part II in 2008. The same form was also used to confirm the pass/fail score for NBDE Part II. The observed scores are candidates’ actual scores. In order to calculate candidates’ true scores, a mathematical formula called the beta binomial model was used. It is not possible to ask candidates to take multiple versions of a licensure examination; therefore, mathematical formulas were used to estimate
true scores. After true scores are estimated, one can evaluate classification accuracy of observed scores.

The classification accuracy with the probability of correct classification, false positive rate, and false negative rate are shown in Table 2. The sum of values of correct classification, false positive, and false negative is equal to 1. As shown, the probability of correct classifications is very high, i.e., 0.97 for the NBDE Part II. Table 2 also shows that the values of false positive and false negative are very small for the NBDE Part II.

**Discussion**

There are three important follow-up issues related to the classification accuracy results reported here. First, additional analysis of the 2009 and 2010 examinations supported the findings from the 2008 exam. Second, the 95 percent confidence interval is the most appropriate framework for analyzing classification accuracy. Third, the beta binomial model produced estimated true scores that closely reflected the actual scores.

The classification accuracy of the new standard was first applied to scores for the 2007 examination. To further validate the new standard, the beta binomial model was also applied to scores from the 2009 and 2010 examinations. Results were quite similar to the 2008 examination scores. The new standard correctly classified 96 percent of candidates in 2009 and 97 percent of candidates in 2010. The failure rate for the reference group taking the 2010 examination is 10 percent. This percentage is significantly lower than the 13.7 percent failure rate in 2009. We do not know what caused the decreased failure rate in 2010. We reviewed potential causes related to the exam: procedures used to set the pass/fail score, changes in the cognitive level of questions (recall vs. application of knowledge), and procedures used by various test administration sites based on the post-exam survey feedback from candidates. None of these factors can account for the decreased failure rate in 2010.

A second issue is using the 95 percent confidence interval to set the pass/fail score for the National Board Dental Examination. Some researchers, practitioners, or educators in dentistry might argue for conducting classification accuracy using 85 percent or 90 percent confidence intervals. Using the 85 or 90 percent confidence internal might result in a larger proportion of estimated classification errors than the .03 found in this study. The 95 percent confidence interval has historically been used by the National Board Dental Examination, and we are reluctant to change it.

A third issue is the accuracy of the beta binomial model to estimate candidate true scores. Review of the predicted true scores showed that differences from the observed scores were negligible; therefore, the observed scores fit the predicted scores well. The model-fit statistics are essential as a validity index for using the beta binomial model to estimate examinee

---

**Table 1. Concept of classification accuracy**

<table>
<thead>
<tr>
<th>True Status Based on All Forms Average Score (True Score)</th>
<th>Decision Made for the Form Actually Taken (Observed Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pass</td>
</tr>
<tr>
<td>Pass</td>
<td>Correct Classification</td>
</tr>
<tr>
<td>Fail</td>
<td>False Positive</td>
</tr>
</tbody>
</table>

**Table 2. Actual classification accuracy**

<table>
<thead>
<tr>
<th>True Status Based on All Forms Average Score (True Score)</th>
<th>Decision Made for the Form Actually Taken (Observed Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pass</td>
</tr>
<tr>
<td>Pass</td>
<td>Correct Classification</td>
</tr>
<tr>
<td></td>
<td>(0.97)</td>
</tr>
<tr>
<td>Fail</td>
<td>False Positive</td>
</tr>
<tr>
<td></td>
<td>(&lt;0.001)</td>
</tr>
</tbody>
</table>
true scores. The acceptable level of model-fit statistics is critically important to estimating candidates’ underlying true ability levels.

**Conclusions**

When scores on an examination are used as a basis for making high stakes pass/fail decisions, it is necessary to confirm the standard or cut score that separates passing and failing candidates. It is also critical to ensure that the pass/fail point on the examination’s scale is reliable and accurate. This article described the procedure for setting the pass/fail score on the NBDE Part II. Analysis on applying the 2009 standard to examination scores for 2008, 2009, and 2010 suggests that the new standard correctly classifies at least 96 percent of candidates. The classification accuracy of the pass/fail standard supports the conclusion that the pass/fail point on NBDE Part II is a valid guide for making decisions about candidates for licensure.

**REFERENCES**