

WREB 2008 TECHNICAL REPORT

An Evaluation of the
Western Regional Examining Board's
2008 Dental and Dental Hygiene
Examination Program

February 2010

Table of Contents

Part I: Background and Incentive for Testing Program Evaluations.....	1
Part II: Description of the Examination Programs.....	2
Part III: Validity	3
Part IV: Standards for Educational and Psychological Testing.....	7
Part V: Legal Defensibility.....	9
Part VI: Validity Evidence	9
Content-related validity evidence.....	9
Items and rating scales.....	10
Reliability	14
Scaling and Comparability	21
Standard setting	23
Administration.....	24
Scoring	25
Part VII: Summative Evaluation	26
References	27
Appendix.....	28

Part I: Background and Incentive for Testing Program Evaluations

Examining organizations exist to provide one source of information to state licensing entities that decide which candidates receive licenses to practice a profession. Professions that are tested prior to licensure or certification include dentists, dental hygienists, certified public accountants, physicians, teachers, and social workers among many other professions.

A major goal of any dental licensing organization is to increase the likelihood that professionally licensed persons will treat patients safely. The tests administered by examining organizations are intended to provide scores that will allow identification of practitioners who might threaten patient safety if they were allowed to practice in a profession. No test or battery of tests is totally accurate for this purpose, and no system of making pass/fail decisions is infallible. Although aware of the possibility that tests could misclassify candidates, states and other jurisdictions use test results to support their decision to license candidates. Testing specialists have developed a system of validation to support the use of test results for these decisions. This system includes: (1) a logical argument, (2) a claim for validity, and (3) evidence to support the contention that using such test scores to make licensing decisions is fair and legally defensible. Passing tests is not usually the sole requirement for licensure, but in most states and jurisdictions, passing tests is one of the requirements. The Western Regional Examining Board (WREB) develops and administers tests to satisfy states' requirement to evaluate the clinical abilities of dental and dental hygiene licensure candidates.

The WREB Board of Directors (BOD) and committees provide vital functions for the operation of this organization. The BOD establishes policy and approves examination changes. The BOD meets twice a year and when required for special decisions.

WREB's Dental Examination Review Committee (ERC) and Dental Hygiene ERC evaluate the work of sub committees and recommend examination changes to the BOD.

The WREB Executive Committee, which consists of the current officers of WREB, the immediate past President of WREB, and the dental and dental hygiene ERC chairpersons, provides corporate direction and makes decisions that affect daily operations, when BOD decisions are not required.

The Examiner Performance Committees for dental and dental hygiene (this is the WREB Executive Committee for dental) review examiner performance and determine the membership of the examiner pool. Examiners from the examiner pool are selected to participate as members of the teams that score candidate performance at examination sites.

WREB annual technical reports describe testing results and provide validity evidence that supports the use of WREB test scores for licensing decisions. WREB continually introduces improvements. Documents archived in WREB's office attest to continued efforts to improve WREB's examination programs. WREB reviews written comments from candidates, WREB examiners, and dental school faculty. WREB completes annual evaluations of its examination program except for years when external evaluations are commissioned to evaluate the year's testing. The evaluations are based on

recommendations in AADE's *Guidance for Clinical Licensing Examinations in Dentistry* (AADE, 2005) and the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 1999). Presently, the evaluations and reports are similar to the example provided by Dr. Tom Haladyna's evaluation of WREB's 2004 testing. The results of these evaluations and study results are presented to WREB committees for their consideration, action, and recommendations.

Part II: Description of the Examination Programs

The Examination Programs provide test scores to states for use in making licensing decisions for dentists and dental hygienists. The following provides highlights of the programs.

The dental examination consists of four parts:

1. Operative:	48 points
2. Periodontal: a.) Treatment:	8 points
b.) Assessment/Diagnosis:	8 points
3. Endodontics:	16 points
4. Prosthodontics:	8 points
5. Patient Assessment and Treatment Planning:	12 points

The total examination points available are: 100 points.

The minimum passing score (cut score) is 75 points. In addition at least 55% of the possible section points must be earned in each of the four sections. For example, a candidate who earned a total of 80 points would not pass the test if only 10 points had been earned in periodontics. If a candidate scores at least 75 overall, but scores below 55% in a section, the candidate may repeat only that section. The scoring requirement for that repeated section will then be 75% of the points possible.

The dental hygiene examination consists of three parts:

1. Probe Depths/Recession:	15 points
2. Extra/Intraoral Exam:	10 points
3. Calculus Removal and Tissue Trauma:	75 points

Total examination points available are: 100 points.

The minimum passing score (cut score) is 75 points.

The dental hygiene restorative examination consists of two parts:

1. Place, carve and finish one composite restoration: 50 points
2. Place, carve and finish one amalgam restoration: 50 points

Total examination points available are: 100 points.

The minimum passing score (cut score) is 75 points.

For all of the testing itemized above, the points available are not raw-score points. Test performance is determined from ratings or validated errors and penalties which are transformed into points using conversion charts that WREB has studied and were ERC and BOD approved.

The dental hygiene anesthesia examination consists of two parts:

1. Written examination - 100 points
2. Clinical examination - Pass/Fail

The minimum passing score (cut score) for the anesthesia written examination is 75 points. The clinical examination is administered only after successful completion of the written examination and is scored by two examiners as a pass or fail based on successful completion of the required procedures.

Detailed information about the examinations can be found in the candidate guides. The candidate guides and other information are on the WREB website: <http://www.wreb.org>

Information about validity can be obtained from technical reports completed each year. This report provides a comprehensive, documented source of validity evidence. Other documents cited in this evaluation also provide validity evidence. WREB has additional documents in the corporate office that provide validity evidence.

Part III: Validity

The most important concern in any examination program is validity. In a high-stakes examination program such as this one, according to leading test expert, Robert Linn (2004), validity takes on more importance due to the fact that a candidate's future as a practicing dentist depends on the outcome of this examination. The test is intended to identify those candidates who are most likely to have a negative impact on public welfare and safety. The focus of this report is validity. All other factors, such as examination content, item quality, reliability, standardized administration, fairness, bias, equity, comparability of scores and scales, and the pass/fail standard are subsumed under validity.

Validity applies to a process involving judgment of the reasonableness of an interpretation or use of a test score. What does a test score obtained from WREB's dental clinical examination mean? How

valid is it for a state to make a licensing decision based on this test score? Thus, validity is not a property of a test, so the term “test validity” is inappropriate. Validity, with regard to WREB testing, focuses on the meaningfulness of test scores based on the scores’ use when making licensing decisions. The first section in the *Standards for Educational and Psychological Testing* (subsequently referred to as the *Standards*) discusses validity in detail and includes numerous standards for validity in testing. The *Standards* document was published in 1999 by the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME). Validity is not an all or nothing property. According to the *Standards*, “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests”.

To support the validity of a test score interpretation or use, four components are useful:

1. A statement that identifies what the test will measure and how the measurement results should be interpreted and used;
2. A claim that the measurement results are validly interpreted and used;
3. A collection of positive and negative evidence relating to this statement and claim;
4. A professional judgment that incorporates this statement, claim, and evidence into a summary statement.

For a positive validation, the argument has to be sound and compelling, the claim justified, and the preponderance of evidence in favor of validity. The effect from any negative evidence should not be of consequence.

No examination program achieves perfection in validity. Improving validity is always a goal. This evaluation report presents the argument and claim for validity, and displays evidence. A summative judgment about validity is presented which is based upon the evidence displayed.

Table 1: Validation of WREB's Dental Examination Program	
Measurement Statement	The American Dental Association's Joint Commission on National Dental Examinations administers national dental and dental hygiene board examinations. These tests measure basic science knowledge and professional knowledge that is believed to be essential for competent practice. The WREB clinical dental and dental hygiene examinations are performance tests that are intended to directly measure clinical competence. Both the national and WREB tests derive principally from practice analyses of the professions of dentistry and dental hygiene. They may include many of the same procedures and consequently, represent complementary aspects of candidate ability. WREB's clinical performance tests are the final elements in the licensing process for dentists and dental hygienists. WREB test scores are intended to be used by licensing agencies along with educational requirements, national board test results and other agency requirements to evaluate candidate qualifications for licensure.
Claim About Validity	WREB claims that scores obtained from testing candidates represent clinical competence and can be used with confidence by participating states, along with other criteria, to make licensing decisions.
Evidence Supporting the Argument	This report provides validity evidence of many types that is based on national testing standards. WREB's technical reports and other documents cited in this report offer validity evidence supporting this argument.
Evidence Weakening the Argument	In this report, to the extent possible, all evidence that weakens this argument is displayed. The negative evidence, found in this evaluation, is minimal. Nonetheless, WREB considers threats to validity and takes action to diminish threats. Those proactive actions strengthen WREB's argument for validity.
Summative Judgment	The claim, arguments for the claim, evidence supporting and weakening the arguments, and any lack of evidence are considered. Then a statement is made about the validity of WREB scores as a measure of professional clinical competence for use by participating states when making licensing decisions

Table 1 shows constituent elements of validation. It describes the process of obtaining evidence to support the validity claim. The table also shows the reasoning process used in validation.

Validity Evidence Used in This Evaluation

Part VI of this report provides details about the validity evidence used in this report. The sources of evidence are many and come in different forms. Recommended procedures, documentation, empirical results including statistics comprise the majority of the evidence found in this report. The validation process should be considered as an accumulation of evidence supporting a judgment about the validity claim. This evidence is used in the same manner that a jury weighs evidence and makes a decision that supports either the prosecution claim or the defense claim.

Evidence Weakening the Argument

Two kinds of evidence that weaken validity are construct under-representation (CUR) and construct-irrelevant variance (CIV). “Construct” as used here, is a name for the domain of knowledge, skills, and abilities that comprise dental or dental hygiene competence. This part of the evaluation seeks to uncover evidence that may work against validity. WREB and licensing agencies, that use WREB scores, do not want such evidence to exist, but its detection and correction are important steps in strengthening the overall validity argument.

CUR is present if the definition of the construct (clinical competence) is not synchronous with what the actual test measures. If we used a multiple-choice test of scientific knowledge or a multiple-choice test of professional knowledge, we would not be representing clinical competence adequately. That is why the national board examinations are necessary licensing requirements but are not sufficient. They under-represent the construct of competence. Considering the results of the national board examinations along with clinical test scores, gives licensing agencies complementary pieces of information that provide adequate representation of the construct of clinical competence. Consequently, licensing agencies value using national board examinations with complementary WREB clinical examinations to evaluate candidates for licensure.

Missing Evidence

Invariably, many testing programs will have some gaps in validity evidence. The intent of this evaluation is to identify any gaps and suggest ways to remove the gaps to improve validity.

Summary

The validation process is summarized in Table 1. It states WREB’s argument that using WREB’s clinical examination scores as a measure of clinical competence is valid. Evidence that supports or weakens this claim for validity is sought. Then a summative judgment is made about the validity of using WREB’s test scores. This judgment can be used by licensing agencies to justify their use of the WREB scores. All licensing agencies have a responsibility to the public to use a valid process for granting licenses. Enabling licensing agencies to identify qualified candidates is WREB’s mission.

Part IV: Standards for Educational and Psychological Testing

A large committee of testing experts and other qualified volunteers participated in developing the *Standards*. These guidelines are used in this evaluation and are cited throughout this document. All of the referenced guidelines bear on the overall judgment of validity.

Table 2: Standards Used in This Evaluation	
Chapter 1: Validity. This chapter identifies fundamental concepts and types of validity evidence that appear throughout this evaluation report.	1.1, 1.2, 1.5, 1.6, 1.7, 1.11, 1.12, 1.15,
Chapter 2: Reliability. As a primary type of validity evidence, evidence is sought	2.1, 2.2, 2.10, 2.13, 2.14, 14.15
Chapter 3: Test Development. Performance testing is recognized as having special challenges in validation.	3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.11, 3.13, 3.14, 3.15, 3.17, 3.19, 3.22, 3.23, 3.24
Chapter 4: Scales, Norms, and Score Comparability including Standard Setting.	4.1, 4.2, 4.9, 4.10, 4.19, 4.21, 14.16, 14.17
Chapter 5: Test Administration, Scoring and Reporting	5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.8, 5.9, 5.10, 5.13, 5.15 , 5.16
Chapter 8: The Rights and Responsibilities of Test Takers	8.1, 8.2, 8.7, 8.11
Chapter 14.8: Testing in Employment and Credentialing	14.8, 14.9, 14.10, 14.11, 14.13, 14.14,

Table 2 lists some of the more important guidelines that were used in this evaluation. Of the many categories that appear in that table and throughout this report, several notable omissions exist that deserve special treatment here.

Chapter 6: Documentation. This evaluation report contains documentation available at the WREB home office that is relevant to the validity claim. This chapter has many categories and issues regarding documentation. This annual evaluation report is one source of documentation. WREB keeps additional documents that support validity. Chapter 6 should be used as a guide for documenting validity evidence. This documentation should be viewed as a kind of insurance that can be used to defend against criticism, legal challenges, and inquiries about the quality of WREB’s examinations.

Chapter 7: Fairness. Since this test is used in licensing, the issue of fairness is an important one. The design and administration of the WREB clinical examinations do not, in any respect, violate any standard of fairness discussed in chapter 7. Scoring examiners have no contact with candidates, and see only their patients. The tests are based on performance and are designed to measure professional competency. There is no threat extant by gender, ethnicity, race, disability or other such factors. Standard 7.12 is general and requires that all candidates be treated fairly and equitably in the examination process. Evidence presented throughout this report verifies the fairness of WREB's clinical examinations.

Chapter 9: Linguistic background. This performance test involves patient treatment under simulated natural conditions involving a patient-dentist relationship. No threat due to inadequate linguistic background is perceived. Candidates are primarily trained in the United States and receive degrees from one of the American dental schools. Foreign trained candidates may have difficulty with the English language. These candidates should be treated fairly. Exploring the fail rates of this population and factors that may contribute to their failing status is desirable. Because gender and ethnicity of most of candidates is unknown to WREB, this kind of study is not possible. WREB committees are aware of the potential problem and attempt to use the simplest language possible in all communication with candidates. All test administrators should also be aware of threats to validity arising from a lack of understanding during WREB tests.

Chapter 10: Testing individuals with disabilities. The 2006 *Dental Candidate Guide* and the dental hygiene Policy Guide discuss provisions for testing candidates with disabilities. Most of the guidelines in the *Standards* (AERA, APA, NCME, 1999) deal with testing elementary and secondary school students. A key issue with WREB's candidates is that each person can be individually assessed with regard to disabilities and then any accommodation in the administration of the test is made in a way that does not alter the competence being measured.

Chapter 11. The responsibilities of test users. This category of standards applies to WREB's participating licensing agencies that use test results. In general, the agencies should have access to all information bearing on the validity of using test scores for making pass/fail decisions. This is an agency responsibility. WREB provides all participating states with information that supports participating states' uses of test scores. WREB's dental and dental hygiene candidate guides provide extensive validity documentation. WREB annual reports provide validity evidence.

Other Standards and Guidelines

Many concepts, principles, and procedures of test development and validation used in this evaluation are based on the preceding standards, but also draw from other important sources. A document that relates the *Standards* (AERA, et al., 1999) to dentistry, *Guidance for Clinical Licensure Examinations in Dentistry*, was published by the American Association of Dental Examiners (AADE) (2005). This document references many of the standards identified in Table 2. It specifically refers to standards that relate to the guidance it provides for clinical testing and for the validity evidence recommended.

Part V: Legal Defensibility

In addition to desiring to provide the highest quality examination program possible, WREB wants to avoid successful legal challenges. Such challenges are expensive and may lead to loss of credibility. Validation is an effort to provide evidence that supports the examination program and its purpose. By assembling validation evidence, WREB provides assurance to its users that WREB test scores are appropriate to use when making licensing decisions. Such validation efforts can also be used with various constituencies and the public to ward off threats of litigation. Potential litigants are disarmed when they know that validation evidence is available. By engaging in evaluations where validity evidence is collected and organized, WREB reduces the threat of legal action. See Mehrens and Popham (1992) for a discussion of legal threats and validity.

Part VI: Validity Evidence

Part VI is the largest part in this report. This part contains a body of evidence intended to support WREB's claim for validity of the use of WREB test scores for licensing decisions. Toward that end, many references to documents are provided in this section. The importance of these references can be found in the *Standards* (AERA, et al., 1999) in chapter 6. Chapter 6 argues that all validity evidence should be documented. In this part of the report, each category of validity evidence is presented. At the end of each category, a brief summary is given and conclusions are drawn about the adequacy of the evidence and the adherence to standards.

The categories are:

- Content-related validity evidence
 - Items and rating scales
 - Reliability
 - Comparability
 - Standard setting
 - Administration
 - Scoring

1. Content-related Validity Evidence

The most fundamental type of validity evidence for a credentialing examination is content. A clinical examination should focus on the skills and abilities needed for successful clinical practice. This domain of skills and abilities is expected to represent professional clinical competence. This category of validity evidence directly addresses WREB's claim for evidence supporting the validity of using this test as a measure of clinical competence. Content-related validity evidence as discussed in the *Standards* can be summarized in this way:

Often a thorough analysis is conducted of the work performed by people in the profession or occupation to document the tasks and abilities that are essential to practice. A wide variety of

empirical approaches is used, including delineation, critical incidence techniques, job analysis, training needs assessments, or practice studies and surveys of practicing professionals. Panels of respected experts in the field often work in collaboration with qualified specialists in testing to define test specifications, including knowledge and skills needed for safe, effective performance, and an appropriate way of assessing that performance (AERA, et al., 1999, p. 156).

WREB committees of subject matter experts, with testing specialist consultation, have conducted practice surveys and practice analyses to identify content and develop test specifications that provide a framework for WREB testing. The development of test items has been a continual process for WREB. Dr. Haladyna has traced this development in his evaluations of the WREB dental and dental hygiene testing programs. This item development history is available in his most recent reports, which are available at www.wreb.org.

2. Items and Rating Scales

WREB's testing programs are complex. Candidates complete performance items where the results are evaluated by dentists and dental hygienists who are trained and calibrated as examiners to use numerous rating scales. To obtain the reliability desired for a high-stakes test where pass/fail decisions are made, it would be desirable for the test items to be numerous, have moderate difficulty for the population tested, and have high discrimination. The selection of the items to be included in the tests should be based on their content and a reasoned analysis of the kinds of tasks in each content area that are linked to clinical competence.

Because WREB tests measure clinical competence in a profession, item scoring is criterion-based instead of norm-based. Many of the items included in the tests are necessary and desirable because the subject-matter committees have determined that these items represent important skills that comprise clinical competence. Consequently, many of the items are included in the tests in spite of low difficulty or discrimination. This has a negative influence on estimates of reliability. Candidates are expected to perform many of the test items (tasks) at a highly successful level. This is true because successful performance is routine when most candidates are very competent. Failure to perform adequately should be rare and would indicate a lack of clinical competence. In a criterion-referenced system with mostly competent examinees, a normal distribution of performance is not expected. As the consequence, the estimated performance characteristics of items may sometimes indicate marginal item functionality when actual item functionality is good. Computed reliability values will generally be lower than true test reliability.

Traditional approaches to item evaluation do not easily apply. This section applies well-established methods and standards to the evaluation of items and rating scales used by WREB with the knowledge that standard evaluative methods are likely to underestimate reliability and other measures of test quality. When performing post test analyses many test performance parameters will be calculated using scores only from candidates with non zero total scores. Candidates are assigned a total score of zero when they test on only one section and also are assigned a zero score when they fail to obtain a qualified patient. To include scores of these candidates would artificially improve reliability estimates and lower candidate examination performance averages.

Dental Hygiene Periodontal Measurements: Probing and Gingival Recession

In a conventional analysis of test item performance, difficulty and discrimination are often computed for each scorable unit (test item). In this performance test, candidates make observations as described in the *Dental Hygiene Examination 2008 Candidate Guide* (WREB, 2008b). The test items, in this instance, evaluate the candidate's skill in measuring probe depths and recession depths. These scores have a low, positive correlation (0.23) with total test scores. The high average score of 14.02 out of 15 possible points in probe and recession limits variability and correlation values. This low correlation limits the usefulness of item difficulty and discrimination values. Validity evidence for this scorable part of the test is based on the judgment, of the committee of experienced dental hygienists, that using these items results in testing that is representative of what dental hygienists are required to do in practice.

Dental Hygiene Extra/intraoral

The nine categories within the extra/intraoral section were identified in the practice analysis as essential aspects of dental hygiene competence. There was essentially no correlation between the individual scores for the first eight categories and total scores. The correlations were less than $r = 0.15$. Their correlations with the extra/intraoral score were low, ranging from 0.33 to 0.41. Low correlations are likely to be the result of the very high scores and the associated low variability. The correlation of classification of periodontal disease with the extra/intraoral score was high at 0.79. The correlation of classification of periodontal disease with the total score was very low at 0.04. This shows that the required skills being measured by the two parts of the examination may not be similar. The choice of these nine categories is based on the practice analysis reported elsewhere in this evaluation and is also based on the recommendations of the Dental Hygiene Examination Subcommittee (WREB, July 18-19, 1998).

Dental Hygiene Calculus Removal

The calculus removal scores had a very high correlation with total scores. The coefficient was 0.93, indicating that most of the variability in the total scores was due to the calculus scoring. The judged quality of the surfaces as observation opportunities for scoring calculus removal and tissue trauma has been verified by the Dental Hygiene subcommittee (WREB, July 2000; July 18-19, 2001; July 5-7, 2002).

Dental Hygiene Restorative

The dental hygiene restorative test is administered for candidates who wish to place and finish restorative materials in states where dental hygienists are authorized to place and finish restorations. The items on the tests are based upon the practice analysis and specifications detailed in restorative sub-committee minutes (WREB, July 17-18, 1999). The scoring of the test items is based upon criteria recommended by the Dental Hygiene Restorative Committee (WREB, July 14, 1998). The correlations of the 6 rated items with the total score were high with values between 0.8 and 0.9. The correlations between occlusal, proximal and margin ratings were moderate ranging from 0.51 to 0.69.

Dental Hygiene Local Anesthesia

The local anesthesia test is administered for candidates who wish to be licensed or certified in a state where dental hygienists are authorized to administer local anesthesia. The testing includes a written and a clinical component with items that are based upon the *Dental hygiene practice survey* (WREB (September 3, 1996).

Dental

Dental subtests, except for Patient Assessment and Treatment Planning, correlated between 0.29 and 0.38 with each other. This is considered to be a low positive range of correlations. The lowest correlations were between Patient Assessment and Treatment Planning and all other sections. These low correlation values were below 0.12. This indicates that the Patient Assessment and Treatment Planning section may be evaluating different skills than what is being evaluated by the rest of the examination. This was anticipated and was the reason for developing the Patient Assessment and Treatment Planning section.

Dental Operative

The operative sub-test has three performance tasks for preparation (13 points for each of two restorations) and three performance tasks for finish (11 points for each of two restorations). *The Candidate Guide* (WREB,2008) provides an extensive presentation of the operative performance tasks and the scoring criteria used. The properties of these items are summarized within the reliability sections under “Descriptive Statistics for Subscales” in this report. Candidates must choose two of three possible restorations: (1) direct posterior amalgam, (2) direct posterior composite or (3) indirect cast gold. The number of candidates, who selected cast gold in 2008, is insignificant.

The correlation between total operative and total test scores was high at 0.88. Operative preparation and operative finish scores were highly correlated with the total operative scores at 0.92 and 0.82 respectively. The operative correlations show that the dental operative subscale is reasonably internally consistent and satisfactory for component scoring within this testing.

Endodontics

The endodontic subtest has two performance measures: access (3 points) and condensation (5 points) for each of two teeth, one anterior and one posterior. *The Candidate Guide* provides an extensive presentation of the endodontic performance tasks and the scoring criteria used. The properties of these items are summarized within the reliability sections under “Descriptive Statistics for Subscales” in this report.

Correlations of scores of each scored item with their criterion (total endodontic score) are moderately high ranging from 0.61 to 0.75. Thus, the endodontics scale appears to be reasonably

internally consistent and items appear to be working as intended given the observation that there is restriction of range and skewness. As the structure study indicated, parts of each subscore have a distinctive structure and are also internally consistent, due to good inter-examiner consistency. The component parts of the endodontics subscore are not highly correlated but high enough to sustain reasonable reliability. The total endodontic score's correlation with total examination score is 0.53.

Periodontal

Periodontal scoring includes a clinical assessment and diagnosis part that resulted from a combined effort between the Central Regional Testing Service (CRDTS), the Southern Regional Testing Agency (SRTA) and WREB. The organization, called CSW, developed a computer administered, simulated patient test. This test contributes a total of 8 points to the WREB periodontal section and the treatment part also contributes 8 points. The treatment scoring consists of 24 observations scored from 0 to 1. Item analysis was not performed here, due to the unique nature of this test. As noted below, the correlation of periodontal assessment and periodontal treatment is very low (0.12) but the correlation of each scale (periodontal assessment and periodontal treatment) to total periodontal score is more substantial, (0.63 and 0.77 respectively). One reason for this is the fact that periodontal treatment is negatively skewed and extremely leptokurtic (peaked). Thus, conventional ways of evaluating variables, such as this one, are not effective. It is important to note the high degree of examiner consistency found in the treatment scoring observations. This high consistency speaks to validity of these assessments given the criterion-referenced nature of candidate performance.

CSW has compiled a year-end technical report that evaluates decision consistency. The results show that the tests are functioning satisfactorily. This report is available at the WREB website: www.wreb.org.

Prosthodontics

CSW developed a computer administered, simulated prosthodontic model test. This functions as the prosthetic section of the WREB dental examination. CSW has compiled a year-end technical report that evaluates decision consistency. The results show that the tests are functioning satisfactorily. This report is available at the WREB website: www.wreb.org.

Patient Assessment and Treatment Planning

The component scores within Patient Assessment and Treatment Planning correlated low to moderately with the total Patient Assessment and Treatment Planning score except for inclusive scores which correlated high (at 0.73) with total section scores.

Summary and Conclusions

Dental hygiene:

The data presented here pertained to statistical performance of items used to create scores for extra/intraoral examination, probe and recession, and for calculus removal. The choices of these items and the underlying rationale are based upon the *Dental Hygiene Practice Survey*, WREB (September 3, 1996) and many discussions by subject matter experts in committees and board meetings referenced at the end of this report. The documentation of the rationales for the observations used to score is a very important source of evidence for item quality. The very high scoring on the dental hygiene testing limits correlation values and reliability estimates. There is strong evidence for consistency in evaluating performance and the documentation of item development and of the rationale for item selection is extensive.

Dental:

WREB's item development and validation activities have been conducted over a long time by four content committees. Minutes and reports provide a good trail of evidence about the process and substance of their achievements. Data presented in this report provide information on item performance. Standard 3.7 applies to item development. WREB is in compliance with this standard. Standards 3.8 and 3.9 deal with documenting field tests and using item analysis. WREB developed the new fifth section, Patient Assessment and Treatment Planning in compliance with these standards.

3. Reliability

A primary form of validity evidence is **reliability**. Because reliability is a primary type of validity evidence, the *Standards for Educational and Psychological Measurement* (AERA, APA, NCME, 1999) has an entire chapter devoted to the topic of reliability. Theoretically, every test score has a random amount of error, which can be large or small, positive or negative. The size and direction of this error are always unknown. However, reliability can be estimated, and from that estimate it is possible to estimate the error in a true score. These estimates provide a way to estimate the average degree of error in a set of test scores. This is an indication of the degree to which test scores might range randomly due to this error. The average error is called the **standard error of measurement**. The error estimation that is most relevant to WREB is the possible error at the passing score, called the **conditional** standard error of measurement.

Several factors contribute to the **estimate** of reliability **underestimating** true reliability (being lower than actual reliability). This lower estimate of reliability does not mean that test scores are less reliable. Instead we should consider the factors that contribute to the underestimation of true reliability when evaluating the reliability coefficient.

One of these factors is the competency level of the candidates taking the test. Restricted samples, where the majority of candidates perform at or about the same levels, cause severe underestimates

of the reliability coefficient. If the sample included a significant number of low performing candidates, the estimate of reliability would be considerably higher. Licensing tests tend to evaluate high-performing candidates, causing this underestimation error in reliability estimates.

Another factor that limits the magnitude of the internal consistency estimates of reliability coefficients is the extent of correlation among the diverse traits (which comprise clinical competence). In some fields, there is a high degree of correlation that leads to high coefficients. In other fields, the skills and abilities are less related. As these reliability estimates depend on internal connections among the elements that make up the score, a coefficient depends on this interrelatedness. In WREB's instances, this inter-connectedness is not very high. Thus the reliability estimates tend to underestimate true reliability.

Reliability Coefficient

There are many ways to estimate reliability, and some lead to spurious estimates of reliability. The most common method is coefficient alpha, which is a measure of internal consistency. Based on the sample of 2,658 candidates who completed all components of the dental examinations, the coefficient for the total examination, using internal consistency analysis software, is 0.89. For the 1314 dental hygiene candidates, the coefficient is 0.79.

The reason for the high coefficients appears to be based on the large number of discrete observations of candidate performance that range on rating scales from 0- 5 or 0-1. The large number of items coupled with some strong correlations among examiner assigned scores tends to improve the reliability estimate of the total score.

The dental candidate scores have a mean score of 81.52 with a standard deviation of 8.91. For dental hygiene the numbers are 89.91 and 9.67. The distribution of scores is negatively skewed, meaning candidates primarily scored near the high end of the score distribution, with a high degree of kurtosis (17.28 for dental), which indicates a peaked distribution of scores. Figures 1, 2 and 3 show this negative skew. Thus, the reliability estimates may be lower than actual reliability. If the candidate pools contained a significant number of incompetent candidates, the estimate of reliability would be higher.

Descriptive Statistics for Subscores

This section provides basic descriptive statistics for the examination subscales. As shown, most subscore means are on the high side of the scale, indicating high performance, as is expected from looking at the total score means. Most subscore means exhibit little variability, once again reflecting high, consistent performance of the candidates. The internal consistency reliability estimates for these subscales is very high considering the small number of items in some of the subtests and the range of the rating scales. These findings indicate that the subscores have fairly high reliability.

Dental Hygiene Examination
(non zero scores)

N of candidates	1454
Minimum	35
Maximum	100
Median	93.00
Mean	90.56
Std. Error (at 75)	.87
Standard Dev	9.63
Skew	-1.92
Kurtosis	4.99

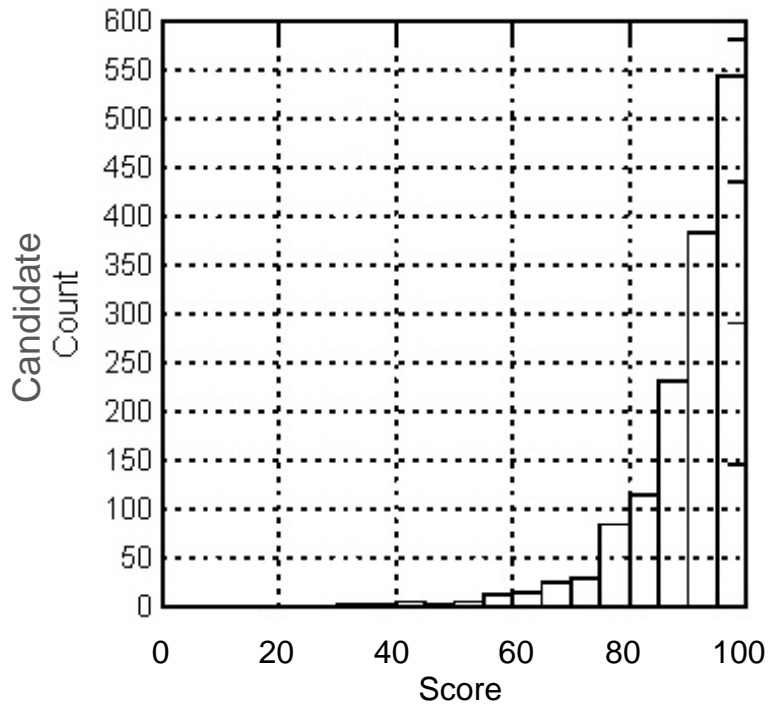


Figure 1.
Dental Hygiene Scores

Dental Hygiene Restorative Examination

N of candidates	301
Minimum	48.5
Maximum	96.1
Median	83.275
Mean	82.136
Standard Dev	6.833
Skew	-0.793
Kurtosis	2.056

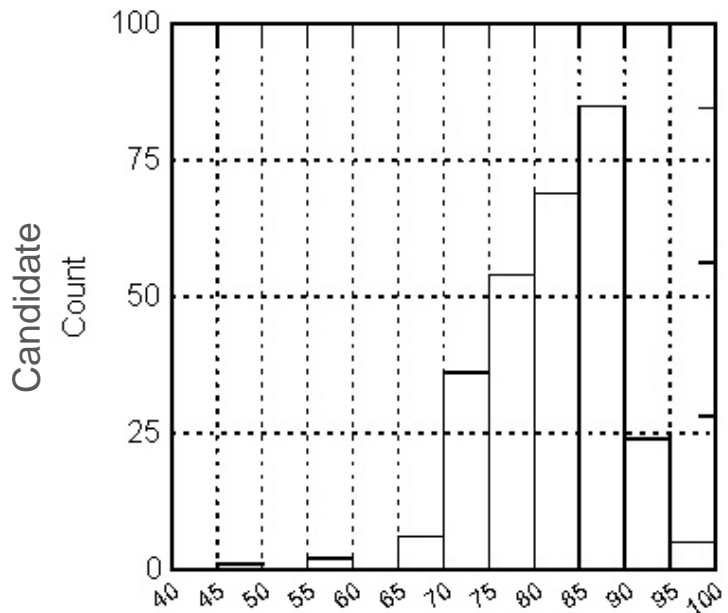


Figure 2.
Dental Hygiene Restorative Scores

Dental Examination (non zero scores)

N of candidates	2371
Minimum	.001
Maximum	96.4
Median	84.23
Mean	82.79
Std. Error (at 75)	.493
Standard Dev	8.631
Skew	-4.36
Kurtosis	33.766

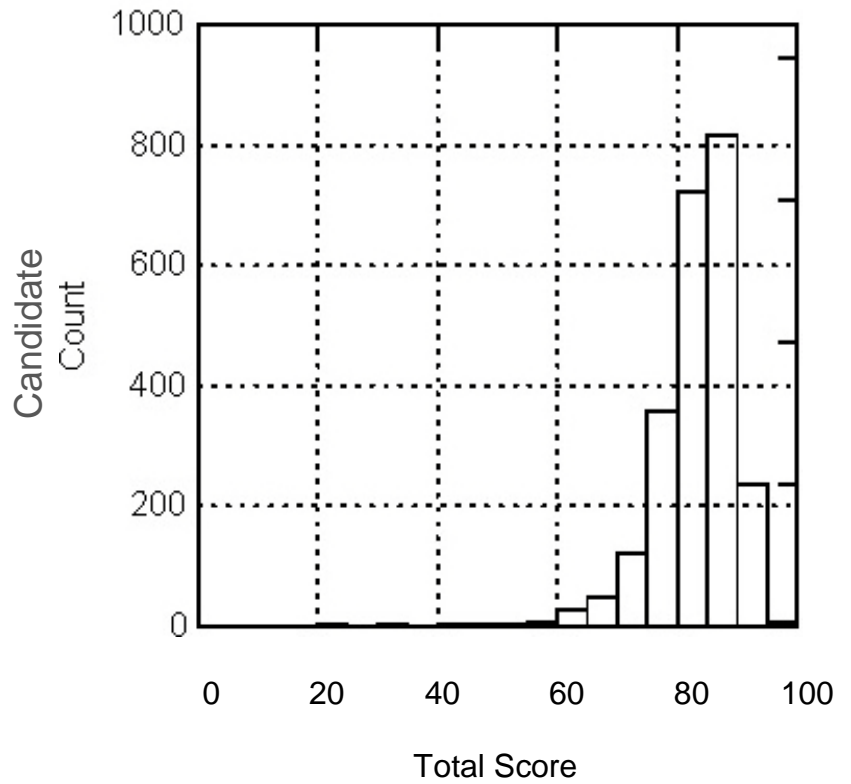


Figure 3.
Dental Scores

Dental Hygiene Pass/Fail Values

Count	Percent	
133	8.9	Fail
1366	91.1	Pass
1499		Total Candidates

Dental Hygiene Subscore Descriptive Values (all candidates with a non-zero total score)

	Extra/Intraoral	Calculus	Probe/Recession
N of candidates	1454	1454	1454
Median	9.35	75.0	15.0
Mean	8.96	69.24	14.0
Standard Dev	1.24	8.88	1.73

Dental Hygiene Testing (Internal Consistency) Reliability

All raw Examiner scores for all candidates with a non-zero total score:
All scores (Cronbach's alpha on 171 examiner scores) = .799

Cronbach's alpha (on 63 examiner scores – does not include probe and recession) = 0.807

Extra intra oral + occlusion + periodontal assessment: Cronbach's alpha (on 27 scores) = 0.832

Calculus: Cronbach's alpha (on 36 scores) = 0.816

Probe and Recession: Cronbach's alpha (on 108 scores) = 0.831

Dental Hygiene Local Anesthesia Pass/Fail Values

Count	Percent	
223	19.5	Failed Candidates
918	80.5	Passing Candidates
1141		Total Candidates

Count	Percent	
391	29.9	Failed Examinations
918	70.1	Examinations Passed
1309		Total Examinations including retakes at same examination site

Dental Hygiene Local Anesthesia Written (Internal Consistency) Reliability

Cronbach's alpha for Anesthesia Written Form A / Form B (on 50 scores) = 0.70 / 0.727

Dental Hygiene Restorative Pass/Fail Values

Count	Percent	
34	11.3	Fail
267	88.7	Pass
301		Total

Dental Hygiene Restorative Testing (Internal Consistency) Reliability

Cronbach's alpha (on 18 scores) = 0.894

Dental Pass/Fail Values

Count	Percent	
332	13.4	Fail
2138	86.6	Pass
2470		Total Candidates

Dental Subscore Descriptive Values

Candidates with non-zero total scores (2371 candidates)

	Prosthodontics Maximum =8	Periodontal Maximum =16	Patient Assessment / Treatment Plan Maximum =12	Endodontics Maximum =18	Operative Maximum =48
Minimum	0.0	0.0	3.57	0.0	0.0
Maximum	8	16.00	12	16.0	48.00
Median	6.92	14.83	10.15	13.69	39.54
Mean	6.86	14.52	9.98	13.13	38.57
Std Dev	0.621	1.22	1.39	2.11	5.34

Dental Testing (Internal Consistency) Overall Reliability Estimate (excludes CSW components)

Estimate is based on all raw scores for all candidates who scored on all procedures
Cronbach's alpha (on 108 scores from examiner) = 0.88

Subcategory (Internal Consistency) Reliability Estimates

Operative: Cronbach's alpha (on 36 scores, 2366 cases) = 0.90
Endodontics: Cronbach's alpha (on 12 scores, 2420 cases) = 0.84
Periodontal Treatment: Cronbach's alpha (on 24 scores, 2367 cases) = 0.75

Operative Component Reliability Estimates

Candidates With Non-Zero Operative Preparation Scores:

Operative preparation: Cronbach's alpha (on 18 scores, 2368 candidates) = 0.86

Candidates With Non-zero Operative Finish Scores:

Operative finish: Cronbach's alpha (on 18 scores, 2366 candidates) = 0.87

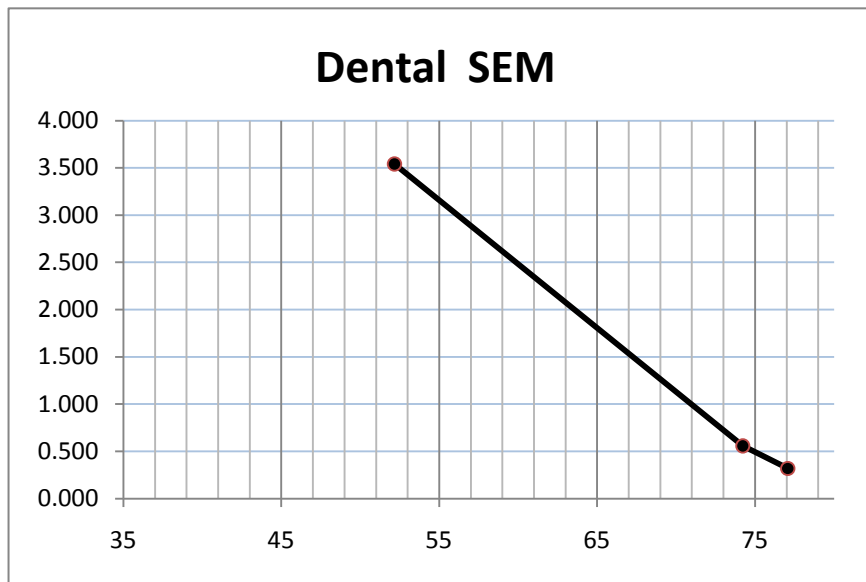
Standard Error of Measurement

The classical standard error of measurement (SEM) provides an estimate of the average test score measurement error for test-takers without regard for their individual proficiency levels. On page 30 of the *Standards for Educational and Psychological Testing*, error near the “cut” score is discussed:

“Where the purpose of measurement is classification, some measurement errors are more serious than others. An individual who is far above or below the value established for pass/fail or for eligibility for a special program can be mismeasured without serious consequences. Mismeasurement of examinees whose true scores are near the cut score is a more serious concern. The techniques used to quantify reliability should recognize these circumstances. This can be done by reporting the conditional standard error in the vicinity of the critical value.”

For a licensing test, it is inappropriate to use the standard error of measurement (SEM) based upon all of the candidate scores. The only goal of licensing tests should be to classify candidates as either competent or incompetent. The passing score which is often called the “cut score” (75 for the WREB tests) is used as the dividing point. Candidates who score below this point are classified as incompetent. It is at this point that accuracy of measurement has a high degree of importance. Consequently, for licensing tests, the local SEM for scores near the passing score is the SEM value of interest. This value is called the conditional standard error of measurement (CSEM).

Agreeing with the *Standards for Educational and Psychological Testing*, the Student Assessment Division of the Texas Education Agency writes: “... it is generally accepted (e.g., Peterson, Kolen, and Hoover, 1989) that the SEM varies across the range of student proficiencies and that individual score levels on any particular test could potentially have different degrees of measurement error associated with them.” Several methods of computing CSEMs are found in the literature, but the methods generally require that the tests have items that are scored dichotomously (right/wrong) and test scores that relate to the number of items scored. In the WREB tests the score is based on the degree of ability on the items, which are weighted differently. Consequently, the available ways of computing CSEM are not appropriate.



$$\text{CSEM} = \text{SEM} @ 75 = 0.493$$

Figure 4.
Dental CSEM as an example of the method used to estimate CSEM

Group	No of cand	Range from	Range to	Mid Pt	Dental SEM	Subsequent groups are not shown here since they are not near the passing score decision point and therefore have little significance with regard to licensing decisions.
1	137	31.944	72.376	52.16	3.541	
2	135	72.362	76.083	74.2225	0.558	
3	136	76.094	78.05	77.072	0.319	

Arizona State University professor/testing specialist Dr. Tom Haladyna suggested that we order the scores from lowest to highest and then divide the group of scores into equal sized groups. The SEM for each group could be found and plotted over the range of scores. From this curve of SEMs, dependent upon score level, the CSEM for the passing score can be found. This results in a passing score CSEM of 0.87 for the dental hygiene test and of 0.493 (CSEM plot shown as an example in figure 4) for the dental test. There were few scores in a skewed score distribution near the passing score in the dental hygiene restorative test. This made an estimate of the CSEM implausible. The reliability estimate for the dental hygiene restorative test was 0.87 for 240 candidates on 18 examiner-assigned scores.

All tests have a margin of error at the passing score. WREB's generally small passing score margins of error may be due to high inter-examiner consistency and high reliability.

Inter-examiner Consistency.

A factor that contributes to reliability is the degree to which examiners rate consistently. WREB has assembled extensive tables and graphs showing examiner consistency. Rather than consider average ratings assigned by examiners, which do not reflect the differences in the abilities of candidates scored, WREB looks at each score assigned by every examiner compared to the average of the scores assigned by the other two examiners for the same scored observation. For scoring using the six-point scale, the percent of times that an examiner agreed with (within 1 point) the average of the other two examiners scoring was 95% for dental hygiene extra/intraoral, 91% for dental hygiene restorative and 90% for dental examiners. For the scores, such as determination of post periodontal treatment calculus remaining, where one or both of the other examiners validated the examiner decision, the median percent was 93% for dental hygiene and 95% for dental examiners.

Summary and Conclusion

WREB meets the applicable standards stated in chapter 2 of *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Standard 2.7 addresses multi-factor tests and suggests that reliability be estimated for each factor. This was done in this report. Standard 2.10 brings up the issue of subjective judging and examiner consistency. The high levels of examiner agreement and score reliability shown here suggest that WREB meets these requirements.

4. Scaling & Comparability

This section addresses the important issue of scaling to achieve comparability of results. A standardized performance test should be consistent from site to site and over years that the examination is administered, so that the cut score is also consistently and accurately applied. The 100-point scale should retain the same meaning each time the examination is given as the difficulty of the examination is the same for every administration.

WREB's examinations are standardized. The examination items are the same each time the examinations are administered. The rating scales are the same. Although examiners who score at each administration may vary, all receive the same training and are calibrated at each examination site. Each examiner's scoring history is considered when assigning examiners to test sites. This is done to assemble examiner teams that are neither harsh nor lenient. The ratings reported in this evaluation show a high degree of accuracy and consistency.

Evidence was presented in the 2003 annual technical report (WREB, 2003a) showing the consistency of results from 1997 through 2003. The appendix of this report shows results since 2003. This shows effective scaling to achieve comparability. The highly standardized nature of this examination, which is discussed in a subsequent section of this report, the examiner consistency discussed in the reliability section, examiner accuracy discussed in the scoring section of this report, and high reliability all provide sources of evidence supporting the integrity of the score scale used in this examination.

Threats to Validity

A threat to validity is a hypothesis that some factor might undermine validity. It is important for WREB to constantly be vigilant to threats and to take action to dismiss or minimize such threats if they prove to be real. In technical testing language, this threat is *construct-irrelevant variance* (CIV). Other synonymous terms are *systematic error* and *bias*. The main idea is that if CIV is detected, it should be eliminated or reduced in seriousness. The rest of this section discusses one of these threats.

Candidates select a patient for their examination. Does the candidates' patient selection render their treatment difficult or easy? Does the condition of the patient in any way affect the outcome of the test? WREB has considered the arguments in favor and against candidates' selecting their patient. WREB has conducted studies to evaluate the extent to which this threat may exist and reported results of a study (Hammond, Fall 2003). The finding was that candidates selecting more challenging patients did not experience lower scores. Thus, CIV due to patient selection did not seem to be a factor in scoring. Critics may use this threat to validity to suggest that WREB's tests are flawed due to this threat. But, the studies completed thus far argue that this threat is not a factor.

Summary and Conclusion

Scaling for comparability is a difficult challenge in high-stakes examinations, particularly where it is performance-oriented and the judgments are subjective. Evidence has been presented in this section and cited elsewhere attesting to the complex issue of scaling for comparability. The candidate guides, the examiner's manuals and WREB's technical reports provide evidence of comparability and how it is achieved. The primary comparability evidence is the reliability and examiner consistency and the extensive training system for examiners. The standardized features of the examinations also contribute.

WREB meets standards for scaling for comparability by creating examinations that are standardized in every aspect, by providing uniform effective training to examiners, and by ensuring that testing conditions are standardized each time an examination is administered. The rating scales provide a scale that is consistently used by examiners. The pass/fail point is built into these scales, and the pass/fail decisions are accurate across all administrations of the tests.

5. Standard Setting

WREB sets its passing scores at 75 and recommends to participating states that its pass/fail recommendations based on performance on the clinical examinations be accepted. Whether a state has a passing standard of 70 or 75 is not relevant. States set arbitrary cut scores as part of their statutes for credentialing examinations. Testing agencies retain the responsibility for setting a fair cut score that meets standards to determine which candidates are recommended for a passing or failing decision. Hammond explains this in an article in the fall 2003 issue of the *WREB Dental Student Newsletter*.

Passing Score Studies

WREB has periodically conducted passing score studies for the written portion of the prosthodontics and anesthesia examinations that are consistent with established standards in the testing industry. For instance, for the prosthodontics test, a passing score study was done using the Ebel method (WREB, June 16-17, 2002). The passing standard for the examiner-rated portions of the examination were incorporated into the rating scales when scoring criteria were developed or revised. As in all other aspects of the testing, the criteria in the rating scales were developed by subject matter experts (with testing specialist consultation) on the subcommittees and were reviewed and approved by the Examination Review Committee and Board of Directors.

Conjunctive Versus Compensatory Standard-Setting Strategies

All licensing boards can elect to use a conjunctive or compensatory strategy in setting a standard. Haladyna and Hess (1999) studied this issue and provided a discussion of both strategies for setting standards. The compensatory strategy allows candidates to make up for a weak performance in one area by a strong performance in another area. The conjunctive strategy requires high performance in all parts of the examination. Most board members in any profession would assert that clinical competence is multifaceted, and evidence of clinical competence should be shown for each important facet. However, making accurate and reliable pass/fail decisions using a conjunctive model is difficult. Most examination programs cannot produce scores that are reliable enough to implement the more desirable conjunctive model. Consequently, many use the compensatory or a hybrid model although the agencies recognize the value of using conjunctive standard-setting strategies.

On February 15, 1999, a memorandum provided background on the issue of conjunctive versus compensatory standard setting. After meetings between WREB officers, CRDTS officers and 5 testing specialists, WREB decided to continue with compensatory scoring because of the high reliability that compensatory scoring provides. Noting that the reliability of the four separate scales

of the current dental examination may now be high enough to support a conjunctive aspect of a hybrid scoring strategy, WREB changed to a hybrid conjunctive/compensatory standard in 2004. This change made a score of 55 percent, of section's points, the minimum in any of the four areas of the dental examination while retaining the compensatory total score standard of 75. The effects of this change were reviewed and little overall adverse impact on candidates was found, but a higher standard than before has been established.

Credentialing boards make adjustments in their standards that are intended to protect the public from unsafe practice. This change represents one of these occasional actions toward that end.

Summary and Conclusion

The technology and methodology for setting cut scores on performance tests, particularly when states have a fixed point in mind, is new and complex. The cut score is mainly based on the examiner judgment as expressed in rating scales. With effective examiner training, a high degree of accuracy and consistency can be achieved which contributes to more accurate scores. Thus, the pass/fail decision is more likely to be accurate under such conditions. Evidence has been presented here and in other sections that address issues like reliability, examiner consistency, examiner accuracy, training of examiners, calibration of examiners, and the creation of many tasks to be rated by examiners. All of these factors have impact on the validity of this cut score.

Based on the evidence appearing in this report and in archives, WREB appears to have substantial support for the validity of making pass/fail decisions using the current cut scores. The rationale for the cut score is well supported and found in the candidate guides, the examiner manuals, and annual technical reports.

Standards 4.19 and 4.21 are related to WREB's examinations. 4.19 asks that the procedures for setting the cut scores are well described and documented. 4.21 states that the manner in which the cut scores are created should utilize expertise of content committees. WREB meets these standards.

6. Administration

Given that the WREB examinations are standardized, the administration of a test must meet certain conditions to provide an equivalent opportunity for success for all candidates. Also, the content of the test must remain exactly the same each time the test is given. And WREB's cut score must be consistently at 75. The candidate guides give a very good account of the many standardized features of this examination. Another important document that provides extensive discussion and information about administration is the *Policy and Procedures Manual* (WREB, 2005). WREB has a staff with complementary abilities that works to achieve a smoothly run examination. The examiner and test administration manuals provide detail on how the tests are administered and scored. The manuals are very detailed, and have evolved over many years. Inspection of the manuals reveals many quality control checks in all aspects of the examination.

As documented in its candidate guides, the examiner manuals, and in other documents in WREB's archive, WREB addresses many issues of administration that affect validity. These issues include training of administrators of the examination, advance information that is available in the candidate guide, clarity of directions in this guide, conditions of testing, patient consent forms, avoiding disruptions in the examination process, test security, monitoring candidates during the examination, responding to questions of candidates, administration instructions, and time limits.

Having a differentiated staff with clear functions is an important aspect of administration. As evidenced in the *Policy and Procedures Manual* (WREB, 2005), WREB has hired and trained staff members who provide valuable service to the administration of the examination. The duties include planning, preparation, administration, and post-test activities. The cycle of activities for each administration is well documented in this manual.

A threat to validity may arise where some test sites are easier or harder than others. Hammond (WREB Fall 2003) discusses this threat and dismisses it with data showing that sites are immaterial as providing an advantage or disadvantage to a candidate. There is no reason or rational hypothesis supporting such a threat. Given the highly standardized nature of this examination, it is unlikely that this threat to validity has an impact on WREB test scores.

Summary and Conclusion

The validity evidence addressing administration is described above and is well documented in the above references. These documents are substantial in scope.

WREB's administration protocols promote consistent, standardized test administration.

7. Scoring

As this test entails clinical performance by candidates for licensing, there are many threats to validity that arise from subjective judgments by examiners. Excellent evidence was presented in the reliability section of this report that attests to high inter-judge consistency in ratings. This section addresses important issues related to examiner scoring.

Selection of Examiners

Examiners are selected and trained as part of the overall administration process. Examiners are licensed professionals in the fields that they are scoring. Documentation of credentials is available for each examiner. The body of information about examiners along with performance data shows the participating states, other constituencies, and the public that WREB has high standards regarding examiner selection and performance.

Training of Examiners

WREB has a training system that has been refined through the years. This training system is well described in the examiner manuals (2007c,d). As specified in the examiner manuals, training is an extensive process that begins with pre-examination training at the examiners' residence or office, followed by a formal exam site training session just prior to each examination. Examiners are sent

a letter, an examiner manual and training materials. These materials are very detailed and provide all examiners with information to help them prepare for scoring candidates.

Examiners are expected to perform adequately. As stated in the examiner manuals, examiner performance is evaluated. Examiners get feedback on their performance and how their scoring varies from their peer examiners. Examiners have to meet an 80% criterion, which means that they can vary from their counterpart examiners less than 20% of the time. If the error rate exceeds 20%, they need to be re-calibrated. WREB has an examiner dismissal clause for examiners (WREB, 2007c,d).

Indications of the extensiveness of this training are also found in content committees' reports. For example, the Operative Committee Minutes (June 21, 2003) provides a discussion of pre-assessment of examiners and details of changes to improve administration. Such discussions are typical of both dental and dental hygiene committee proceedings. All meeting minutes of these committees are archived by WREB, and only a sampling of these has been cited in this report to provide some documentation background.

Summary and Conclusion

The validity evidence addressing examiner scoring is described above and is well documented. The documentation on examiner selection, training, and performance provide an assurance that the scoring process provides candidate scores that can be used for licensing decisions. WREB's examiner protocols promote consistent, reliable test results.

Part VII: Summative Evaluation

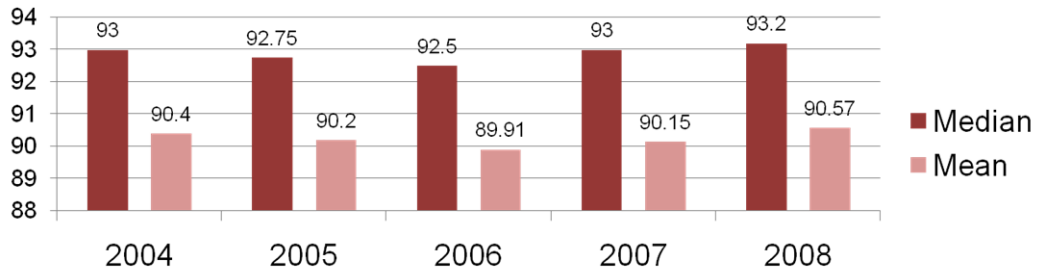
The argument, claim for validity, and evidence presented in this document and in WREB's technical reports, and other documents strongly supports the validity of using test scores for making pass/fail decisions that affect licensing of dentists and dental hygienists in WREB's participating states.

References

- American Association of Dental Examiners (2005). *Guidance for clinical licensure examinations in dentistry*. Chicago: Author.
- American Educational Research Association, American Psychological Association. National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Haladyna, T. M. (2005). An Evaluation of the Western Regional Examining Board's Dental Examination Program. Phoenix: Author.
- Haladyna, T. M. (2005). An Evaluation of the Western Regional Examining Board's Dental Hygiene Examination Program. Phoenix: Author.
- Hammond, D. (Fall, 2003). Obsession with saving the "ideal" lesion for the WREB examination. *WREB Dental Student Newsletter*.
- Linn, R. L. (2000). Assessment and accountability. *Educational Researcher*, 29(2), 4-16.
- Mehrens, W. A., & Popham, W. J. (1992) How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education*, 5(3), 265-283.
- Western Regional Examining Board (September 3, 1996). *Dental hygiene practice Survey*. Phoenix, Author.
- Western Regional Examining Board (July 14, 1998). *Restorative Sub-Committee Recommendations to the DH-ERC*. Phoenix: Author.
- Western Regional Examining Board (July 18-19, 1998). *Clinical sub-committee meeting*. Phoenix: Author.
- Western Regional Examining Board (July 17-18, 1999). *Restorative Committee Meeting Minutes*. Phoenix: Author.
- Western Regional Examining Board (July 2000). Dental hygiene clinical subcommittee recommendations and justifications. Phoenix: Author.
- Western Regional Examining Board (July 18-19, 2001). *Dental hygiene sub-committee meeting*. Phoenix: Author.
- Western Regional Examining Board (July 5-7, 2002). *Report of the dental hygiene sub-committee meeting*. Phoenix: Author.
- Western Regional Examining Board (2005). *Policy and Procedures Manual*. Phoenix: Author.
- Western Regional Examining Board (January 7, 2006). *Western Regional Examining Board By Laws* (As amended by the Membership). Phoenix: Author.
- Western Regional Examining Board (2003) *WREB Annual Technical Report 2003*. Phoenix: Author.
- Western Regional Examining Board (2008a). *Dental Candidate Guide*. Phoenix: Author.
- Western Regional Examining Board (2008b). *Dental Hygiene Candidate Guide*. Phoenix: Author.
- Western Regional Examining Board (2008c). *Dental Examiner Manual*. Phoenix: Author.
- Western Regional Examining Board (2008d). *Dental Hygiene Examiner Manual*. Phoenix: Author.

Appendix

Annual Median and Mean for WREB Dental Hygiene Exam



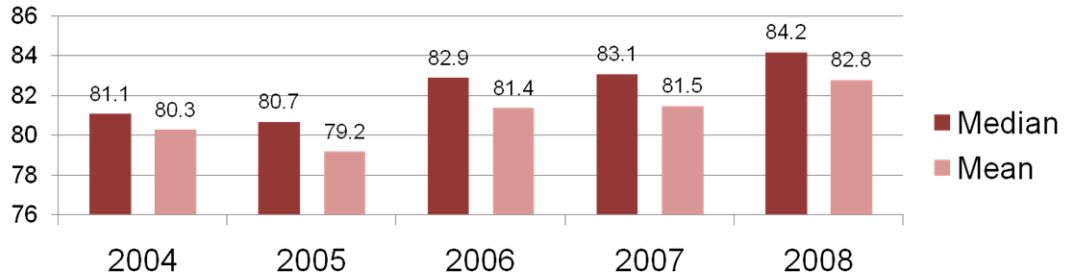
Number of candidates by Year

Year	2004	2005	2006	2007	2008
Number of candidates	1183	1184	1356	1434	1499

Annual Failure Rates for WREB Dental Hygiene Exam



Annual Median and Mean for WREB Dental Exam



Number of candidates by Year

2004	2005	2006	2007	2008
1361	1793	2351	2747	2470

Annual Failure Rates for WREB Dental Exam

