An Evaluation of the

Western Region Examining Board

Dental Examination

Dr. Thomas M. Haladyna
Professor Emeritus
Arizona State University

May 27, 2010
tmh@asu.edu

Acknowledgments

The evaluation of any examination program is complex and challenging. As with the previous evaluations, Del Hammond provided considerable and valuable assistance in explaining the complexities of this examination process and giving me many useful, well organized, and accurate data files and documents that made this report possible. Denise Ramos also provided many documents that aided my review, which appear in the appendix. Radley Masinelli provided some excellent information about WREB's security policy and procedures. Shane Clays of CWS Computer Simulations also provided information about the periodontal subtest and the Prostodontics test. WREB's board is commended for undertaking this evaluation. I am solely responsible for any errors or lack of clarity in this document.

Dr. Thomas M. Haladyna
Phoenix, Arizona
May 2010

Table of Contents

INTRODUCTION

Examining boards like WREB periodically undergo an external evaluation to determine how validly test scores are interpreted and used by participating states for licensure decisions. An external evaluation provides an independent, professional, qualified opinion about an examination program's validity of test score interpretation and uses. The process of evaluation entails many steps explained in the section on validity. The organization of this report is complex because the examination is scored conjunctively, that is, the examination consists of five independently scored tests. Because each test is scored independently, a section is devoted to the validity of test score interpretation and uses for each test.

For the purpose of clarity of language, the term *examination* is used to refer to the entire testing program, which is the *Dental Examination*. As this examination consists of five parts, each part will be referred to as a *test* because each part is a test that stands alone regarding validation.

The 12 sections of the report are briefly summarized below to give the reader an overview of what follows.

I. Explains the reason for this evaluation.
II. Describes the *Dental Examination*.
III. Explains *validity* and the investigative process known as *validation*.
IV. Discusses national testing standards followed in this evaluation.
V. Mentions the threat of legal challenge and how to defend against it.
VI. Covers the complexity of conjunctive scoring and how validity evidence was organized for this evaluation.
VII. Presents validity evidence used for all five tests that comprise the *Dental Examination*.
VIII. Presents unique validity evidence for the *Operative* test.
VIII. Presents unique validity evidence for the *Periodontal* test.
IX. Presents unique validity evidence for the *Endodontics* test.
X. Presents unique validity evidence for the *Prosthodontics* test.
XI. Presents unique validity evidence for the *Patient Assessment and Treatment Planning (PAPT)* test.
XII. Provides a summative evaluation.

PART I: WHY IS THE *DENTAL EXAMINATION* BEING EVALUATED?

WREB is an organization that conducts clinical examinations in dentistry and dental hygiene. This organization was formally incorporated in 1976. WREB has provided services to states and candidates in growing numbers. Its corporate office is in Phoenix, Arizona. Its bylaws were amended by its membership (WREB, January 11, 2003; January 7, 2006a). A history of WREB is available on its website (wreb.org/Information/History.htm). Another good source is any annual report (e.g., WREB, 2008).

Examining boards provide important information to states, who must decide who receives a license to practice a profession in that state's jurisdiction. These professions include dentistry, dental hygiene, accountancy, architecture, medicine, education, social work, law, and law enforcement among many others. WREB provides this service to 14 member states and participating states. Test scores are used with other information to decide licensure for each candidate in a state.

WREB has a Board of Directors (also known as the *governing board*). This board meets twice annually to discuss policy and oversee examination development and validation. Meeting minutes provide a historical documentation of changes in the *Dental Examination* leading up to the transition from a compensatory scoring model to a conjunctive scoring model (WREB, July 5, 2005; January 7, 2006b; July 20, 2006; January 6, 2007; July 12, 2007; January 5, 2008, July 10, 2008; January 10, 2009, July 16, 2009; January 30, 2010).

An Examination Review Committee oversees the *Dental Examination* Program. WREB has a committee for each of the five tests. These committees meet regularly, review policies and procedures, and recommend changes intended to improve the examination program. The structure of committees and the way staff serves WREB and the committees are clearly shown in any annual report (WREB, 2010). The Examination Review Committee also issues annual reports at the Board of Directors meetings (WREB, July, 19, 2006; July 11, 2007; July 9, 2008; July 15, 2009; 2010b). These reports provide information about the transition from the past compensatory-scored *Dental Examination* to the current conjunctively-scored *Dental Examination*.

Responsibilities of Examining Boards Like WREB

The main concern of any examining board is to increase the likelihood that a professionally licensed person will treat their patients safely. The content of these examinations is professional competence. This content usually consists of knowledge, skills, and abilities (KSAs). Defining KSAs is a very important task of these examining boards that affects validity.

No examination or battery of tests is infallible in helping identify candidates who might jeopardize public safety. Nonetheless, all states and jurisdictions engage in licensing examinations to inform decision making about who receives a license to practice a profession. Of course, the examination alone does not determine who receives a license. In most states and jurisdictions, passing an examination is one important criterion for licensure that all candidates must achieve if they are to be allowed to practice in that state.

Although WREB is a testing agency that provides validated test scores, it is each state's responsibility that the interpretation and use of test scores are valid. WREB provides assurances through its reputation but more importantly through its documentation of validity that the test scores and the cut score guidelines it provides are in the best interests of the states and the citizens of each state.

Evaluation of an Examination

An external evaluation of an examination is highly recommended by testing experts (Buckendahl & Plake, 2006; Madaus, 1992; Downing & Haladyna, 1996). The benefit of such an evaluation is to verify that the examination program is providing valid information about the professional competency of its candidates. Such evaluations also provide constructive criticism intended to improve validity.

Every examination program consists of three important, logical, sequential, related elements:

1. defining of a profession as to KSAs needed to practice safely and competently,
2. development of an examination that validly measures competence in the profession, and
3. validation of the interpretation and uses of examination scores derived from administering that examination.

Testing specialists have developed a way of thinking about validation (e.g., Kane, 2006a, Kane, 2006b). One might think of validation as an investigation bearing on validity. This investigation involves many related steps that include an argument about validity, the gathering of evidence to support the argument, and an evaluation of this evidence concerning the argument.

Two earlier evaluations of WREB's *Dental Examination* provided validity evidence and opinions that were current to the date of each evaluation's publication (Haladyna, 1998; 2005). The organization and emphases in the current report differ slightly to reflect changes in the concept of validity and validation (Kane, 2006a, 2006b). Greater emphasis is placed on reliability in this evaluation. Also, as the current examination consists of five independently scored tests, this report will present validity evidence for each of the five tests. This difference is because the previous examination had a single score, whereas the current examination contains five tests, each of which yields pass/fail scores.

PART II: DESCRIPTION OF THE DENTAL EXAMINATION

Table 1 provides highlights of the *Dental Examination.* More detailed descriptions of the five tests of this examination appear in the *Dental Examination Candidate Guide* (WREB, 2009b).

Table 1: Highlights of the *Dental Examination*

Five tests comprise this examination: 1. Operative, 2. Periodontal, 3. Endodontics, 4. Prosthodontics, and 5. Patient Assessment and Treatment Planning. As WREB has adopted a conjunctive scoring model, a successful candidate must pass each test.

For those tests that are appropriate to score by percentage, the score is expressed in a 100-point scale and 75 is the cut score (passing level). Those sections that do not lend themselves to percentage scoring are graded on a rating scale that varies from one to five. The cut (passing) score is three on this scale. The median examiner rating for each criterion category is multiplied by the weight for that category. These weighted ratings (less any deductions) are added to obtain the score for each procedure. The scores are then averaged to obtain the overall section score. If the section average of three were converted to a comparable percentage score, it would correspond to a minimally acceptable level of 75. The two Periodontal sub-tests are weighted equally and are combined for the periodontal score with a cut score of 75.

The Periodontal Assessment/Diagnosis sub-test, the Prosthodontics test, and the Patient Assessment and Treatment Planning test are administered by computer at the Pearson-Vue centers throughout the United States. Another company, CSW Computer Simulations is responsible for the test development and validation of the Periodontal test and the Prosthodontics test.

Information about validity can be obtained from annual technical reports and other documents in the archive and from previous evaluations (Haladyna, 1998, 2005; Hammond & Buckendahl, March 2007; Hammond, February 2010). This evaluation report provides references to many other documents bearing on validity in WREB's archive.

Examiners receive training in the examination process and are validated using a performance examination. These examiners seldom deviate from one another in their judgments. Harshness and leniency in ratings of these examiners were very low. Sources of data bearing on this threat to validity are presented in this report.

More detailed information about this examination program can be found in the *Dental Candidate Guide* (WREB, 2010a). Another source of information is the WREB web page: http://www.wreb.org/

It is the responsibility of each candidate for licensure to identify and bring a qualified patient to the examination site. Each candidate will treat a patient. Highly skilled and well-trained examiners evaluate the results of each candidate's performance. WREB has developed extensive criteria for appointing the examiners (WREB 2009a). Professional judgment is a key element in determining these scoring rules. All judgments are validated, as will be described later in great detail. Performance as determined from ratings is transformed into points using conversion charts that WREB has studied and approved by a committee of its SMEs through a consensus. These

charts also appear in the *Dental Candidate Guide* (WREB, 2009b). Scoring by examiners is mainly objective but rating scales are also used that require subjective judgment. These results are used to compute a total score for each of the five tests comprising this examination. The cut score (passing score) will be described in subsequent appropriate sections.

PART III: VALIDITY

A test score should provide a valid interpretation of a candidate's professional competence. If a test score is used as one criterion to advance or prevent advancement of a candidate to licensure, the decision to pass or fail must be extremely valid. The focus of this evaluation is validity. All other ideas about test quality are subsumed under validity.

Validity involves the professional judgment about the reasonableness of an interpretation or use of a test score. The *Standards for Educational and Psychological Testing*–hereafter referred to as the *Standards* (American Educational Research Association-AERA, American Psychological Association-APA, & National Council on Measurement in Education–NCME, 1999) provides guidelines for evaluating validity. Additionally, the American Association of Dental Examiners–AADE (2005) issued guidelines for clinical performance examinations that include both dentistry and dental hygiene.

What does a test score obtained from the *Dental Examination* battery of five tests mean? How valid is it for a state to make a pass/fail decision based on a score for each of these five tests? Validity focuses on the meaningfulness of an interpretation of a test score and the reasonableness of its use in making pass/fail decisions.

As noted previously, the investigation process for evaluating validity is validation (Kane, 2006a). This process begins with a definition of dentistry that is usually derived from a practice analysis (Raymond & Neustel, 2006). Then to validate interpretations and uses of examination scores, we need certain elements in this validation:

1. an argument that describes what WREB plans to measure and how test scores will be validly interpreted and used;

2. a claim that the test scores are validly interpreted and used;

3. a collection of validity evidence related to this argument and claim; and

4. a professional judgment that incorporates this argument, claim, and evidence into a summary judgment.

For a positive evaluation, the argument has to be sound and compelling, the claim just, and the preponderance of evidence in favor of the stated interpretation and use of test scores. Negative validity evidence should be inconsequential.

No examination program reaches its ultimate in validity. Validity is a goal. All examination programs undergo improvement in an evolutionary path, but the road is steep and long. This evaluation report presents the argument and claim for validity, and it displays the evidence supplied by WREB and CSW Computer Simulations. The author of this report has evaluated the argument and evidence to make a summative judgment about validity of each of the five tests.

Table 2 shows the constituent elements in validation. This table also shows the reasoning process used in this validation.

Table 2: Validation of WREB's *Dental Examination*	
Argument	The American Dental Association administers a *National Board Dental Examination*. This examination measures the knowledge and skills thought to be necessary for safe and competent practice. This examination derives principally from a practice analysis of the profession of dentistry. The WREB *Dental Examination* is a performance examination intended to measure clinical competence. It also is based on practice analysis. These two examination programs represent complementary aspects of dentistry. WREB's *Dental Examination* is the capstone in this licensing process for dentists.
Claim About Validity	WREB claims those examination scores obtained from 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 evaluation report provides validity evidence of many types that are based on national test 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, evidence is displayed that weakens this argument. In the judgment of this evaluator, this kind of evidence as discussed in this report is inconsequential to validity. Nonetheless, WREB should endeavor to consider threats to validity and act accordingly to diminish the threat. By that, WREB strengthens the evidence supporting the argument and the claim for validity.
Lack of Evidence	Gaps in evidence are noted in this report.
Summative Judgment	This evaluator considers the argument, claim, and evidence before making a judgement about validity of WREB scores as (1) a measure of professional clinical competence, and (2) for use by participating states in making pass/fail decisions.

Validity Evidence Used in This Evaluation

Parts VII to XII of this report provides validity evidence. The sources of evidence are information found in documents and statistical analysis. Validity evidence should never be counted in a checklist. Instead, the evaluator considers the body of evidence before making a summative evaluation. This evidence is used in the same manner that a jury considers evidence and decides that it supports either the prosecutor's claim or the defense's claim.

Evidence Weakening the Argument

In any evaluation for validity, honest examination of evidence that undermines validity is seldom done by examination sponsors (Cronbach, 1988). According to Messick (1989), two kinds of evidence that weaken validity are construct under-representation (CUR) and construct-irrelevant variance (CIV). The construct is another name for the domain of KSAs that comprise competence. This part of the evaluation seeks to uncover evidence that may undermine validity. Naturally, WREB and its client states do not want such evidence to be strong, but its detection and eventual treatment are important steps in strengthening the overall validity argument and related claim. Every examination program is only as strong as its weakest link. For most testing programs, a validity research agenda is useful for exploring problems and solving problems that bear on validity (Haladyna, 2002b, 2006).

CUR is present if the tests used to measure competence do not match very well the definition of dentistry. ***Fidelity*** is the technical term we use to assess the connection of the tasks on the examination to the definition of dentistry. If we used a multiple-choice examination of scientific knowledge or a multiple-choice examination of professional knowledge, we would not be representing clinical dental competence adequately. That is why the *National Board Dental Examination* is a necessary licensing requirement but it is not sufficient. These multiple-choice tests under-represent the construct of competence in dentistry. When we combine the results of the *National Board's Dental Examination* with WREB's *Dental Examination*, we have important complementary pieces of information that provide adequate representation of the construct of dentistry. Thus, participating states see the value of using both the National Board's and WREB's examination due to their complementary nature with respect to the KSAs that comprise professional competency.

Summary

This section on validity is best summarized in Table 2. It shows that we start with a definition of dental competence, then formulate an argument about the validity of using WREB's *Dental Examination* test scores as unique, complementary measures of clinical competence. A claim is made by WREB for its client states that using these test scores in that way is valid. Validity evidence is collected and displayed. After all evidence is assessed, a summative judgment is made about the validity of each test. Participating states can use this judgment to guide them in deciding if the test score information they receive is adequate for their needs. As mentioned previously, all licensing authorities have a responsibility to the public to do this. WREB exists to help these states accomplish this mission.

PART IV: *STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING*

The *Standards for Educational and Psychological Testing (*subsequently referred to as the *Standards)* 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). A large, representative committee of testing experts and other qualified volunteers participated in developing these guidelines. For this evaluation, these guidelines are used and often cited throughout this document. All of the referenced guidelines bear on the overall judgment of validity. A new set of standards is being developed, but these standards will not be published until 2012 or later. That is why the current standards are used for this evaluation. The American Association of Dental Examiners (2005) published *Guidance for Clinical Licensure Examinations in Dentistry*. Although not specifically cited, these guidelines also apply to this evaluation. The two sets of guidelines are very similar in content, but the former focuses on clinical examinations like WREB's.

Table 3 on the next page summarizes some more important standards that will be cited and used in this document. 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 ***all*** documentation made available by WREB used for the validity claim stated in this evaluation. This chapter has many categories and issues regarding documentation. The annual technical report is one source of documentation. This evaluation report is another source of documentation. WREB keeps an archive of other documents that bear on validity. These documents are cited in the reference list at the end of this report. Chapter 6 should be used as a guide for documenting its 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. Other sources of information about the importance of documentation include Becker and Pomplum (2006) and Haladyna (2002a).

Chapter 7: Fairness. As this examination is used in licensing dentists, the issue of fairness is an important one. The design and administration of the *Dental Examination* do not in any way violate any standard of fairness discussed in chapter 7. Examiners have no contact with candidates. Examiners only see the candidate's patients. As this examination is based on performance and measures professional competence, no threat extant by gender, ethnicity, race, disability or other factors seems imminent. Standard 7.12 is the most general of these and requires that all candidates be treated fairly and equitably in the examination process. Evidence presented throughout this report bears on the judgment of fairness of the *Dental Examination*.

Chapter 9: Linguistic background. As this performance examination involves patient treatment under simulated natural conditions involving patient-dentist interactions, no threat due to inadequate linguistic background is perceived. Most of the candidates are trained in the United States and received their degree from a U. S. dental school. Foreign trained candidates often have difficulty with the English language. These candidates should also be treated fairly. All examination sponsors should always be alert to any threat arising from a lack of understanding of the recommended procedures for this examination or other factors that may jeopardize a candidate

whose primary language is not English. A subtle point is that language should be appropriate for the practitioner. The language of the examination should not be pedantic or unnecessarily complex linguistically. On the other hand, this examination should not simplify the language inappropriately to adapt to an English language learner. The reason is that every candidate needs to have sufficient verbal ability to read, write, speak, and listen at an appropriate level for the profession of dentistry in the United States.

Table 3: Categories of 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: Examination 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: Examination 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 Examination 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

Chapter 10: Testing individuals with disabilities. Page 3 of the *Dental Candidate Guide* (WREB, 2009b) discusses provisions for testing candidates with disabilities. Most of the guidelines in the *Standards* (AERA, *et al*., 1999) deal with testing elementary and secondary school students. A key issue with WREB's candidates is that each person is individually assessed regarding disability and then any accommodation in the administration of the test is done in a way that does not alter the competence being measured.

Chapter 11. The responsibilities of test users. Overall, participating states should have access to all information bearing on the validity of using examination scores for making pass/fail decisions. This is a state's responsibility; it is not WREB's responsibility. However, WREB is responsible for giving all participating states such information that supports their uses of examination scores. WREB's *Dental Candidate Guide* (WREB, 2009b) is published every year and provides much information. WREB's technical reports are sources of information (Hammond & Buckendahl, March 2007; Hammond, February 2010). These reports are sources of information that are available to states. WREB's website provides public access to these documents.

PART V: LEGAL DEFENSIBILITY

No examining board wants to be challenged legally for adverse test score decisions that might be considered invalid. Such challenges are expensive to defend and if successful may lead to loss of credibility that can ultimately weaken and destroy an examination program.

Validation provides evidence that supports the examination program and its purpose. By undertaking a validation, WREB provides assurance to its participating states that the examination score information can be used validly. Therefore, validation can discourage unwarranted litigation. When potential litigants know that validation has been done and the validity evidence is publicly available, they are less likely to challenge the examining board's test score interpretations and uses.

In all circumstances, any examining board should have continued legal counsel that examines threats that arise from legal actions and its position in thwarting these threats. By engaging in this evaluation where validity evidence is collected and organized, WREB very effectively reduces the threat of legal action. Mehrens and Popham (1992) provided a useful discussion of legal threats and validity. By paying particular attention to validity, the intent is often sufficient to ward off legal challenge.

WREB has made public evidence in technical reports and evaluations such as this one. WREB's website is very informative and represents a model for other examining boards. (See http://www.wreb.org/Information/Articles.aspx). WREB's annual reports provide useful overviews of its examination program and other sources of information about its programs (WREB, 2009).

PART VI: CONJUNCTIVE SCORING AND ITS IMPLICATIONS FOR VALIDITY AND VALIDATION

Where a battery of tests is given and a total score is computed based on the sum of these test scores to make a pass/fail decision, the scoring model is *compensatory*. That is, a low score in one test may be inconsequential if the other test scores are high enough to warrant a passing decision. A compensatory scoring model involves total overall performance. The argument is that a low score on one test should **not** result in failure, if other performances are higher and, therefore, compensate for a low performance. If the total score meets or surpasses the cut score, the candidate passes.

Where the sponsors of a credentialing testing program believe that a low score in any test or subtest may infer that the candidate's professional practice may harm a patient, a conjunctive scoring model is preferred to a compensatory scoring model. Each test in the examination is subject to a pass/fail decision, and a candidate cannot meet a state's standards until all tests in the examination are passed.

Haladyna and Hess (1999) provided a comprehensive discussion of the pros and cons of each scoring model and the responsibilities of testing agencies in validating test score interpretation and uses. The decision about which scoring model is appropriate is best made by the governing board.

The compensatory scoring model sets a lower standard for passing and failing. Because it combines scores from related tests, reliability tends to be high. If competency is viewed as a holistic concept, the compensatory model is appropriate.

The conjunctive scoring model is more restrictive and sets a higher standard for passing. Each test in the battery is subject to validation. Each test must measure a distinctly different aspect of competence. Each test should yield sufficiently high reliability coefficients, so that pass/fail decisions can be made with some confidence. Typically where patient or client welfare is at stake, conjunctive scoring models are used. For instance, some professional testing programs where conjunctive scoring is used are: the American Institute of Certified Professional Accounts, Registered Architects, and Arizona Peace Officers. Candidates for licensure must pass all tests to qualify for licensure. When a candidate fails a test in the battery, the candidate is given opportunities to retake the test, often after a reasonable waiting period and remediation.

WREB has determined that the licensing of dentists involves five important areas of competency and a pass/fail decision is justified for each of these areas. The history of this important transition is well documented in the Examination Review Committee Minutes (July 11, 2007; July 9, 2008; July 15, 2009). At the July 11, 2007 meeting, the issue of conjunctive versus compensatory scoring was discussed. At the July 9, 2008 meeting, the motion to use a conjunctive scoring system was passed. As a result, WREB instituted conjunctive scoring for the examination in 2009. This revised scoring was for five independently scored tests where pass/fail decisions were made on each of the five tests. Several types of validity evidence contribute to the wisdom of such a decision. The most important of these considerations is that content areas can be identified and prioritized via a practice analysis. These five content areas must be clearly identified as uniquely important to patient safety. Another consideration for a conjunctive scoring model is that

the reliability of each of the test scores is sufficiently high to warrant an accurate pass/fail decision. Generally, we like these coefficients to be above 0.80. However, a proviso that is necessary here is that with tests where competency is very high, test scores tend to be negatively skewed. Technically and statistically, this narrow distribution of scores is known as a *restriction in range*. As a result, the estimation of reliability is often low, not because the test scores are unreliable but because the restriction in the range of scores affects the reliability estimation. The defense against misinterpretation of such results is to report the standard error of measurement, which most often will be very small. With a small standard error of measurement, the uncertainty of the few candidates whose score fall at or near the cut score (passing score) is minimal.

In the remainder of this report, the Part VII is devoted to validity evidence that applies to all five tests, and then, Part VIII to Part XII are devoted to presenting validity evidence that is unique to each test.

PART VII: VALIDITY EVIDENCE THAT APPLIES TO ALL FIVE TESTS

Content-related Validity Evidence

The most fundamental basis for identifying the content of any professional credentialing examination such as this one is to conduct a practice analysis (Raymond & Neustel, 2006). This survey of the profession provides information about the knowledge, skills, and abilities (KSAs) needed to practice competently and safely in WREB member states. A practice analysis was completed (WREB, May 15, 2007), which formed the basis for the new 2009 *Dental Examination.* In subsequent sections, the basis for content is presented. All these test development activities for the five tests comprising this examination arise from the content definitions provided by the 2007 practice analysis.

An issue facing all test developers for a clinically-based professional competency examination is whether each test has ***fidelity*** with its criterion. Fidelity is a judged characteristic of any test by which each test item should resemble or replicate what a practicing professional does in actual clinical practice. Testing agencies need to show how much each test is relevant to practice as shown in the practice analysis. Typically, such tests have little fidelity because developing tests that directly measure actual practice procedures is very difficult. The clinical examination in dentistry is a rare instance of a test with extremely high fidelity. The tasks performed by candidates on the clinical examination resemble those done in actual practice. This statement applies to the *Operative* test, the *Endodontics* test, and one part of the *Periodontal Test*. The subtest of the *Periodontal Test* known as *Periodontal Assessment/Diagnosis* is a high fidelity simulation, although it does not deal with an actual patient. It has good fidelity, but not as high as a test where the actual skills are performed. The *Patient Assessment and Treatment Planning* test is also a high-fidelity simulation. Finally, the *Prosthodontics test* is a simulation where MC items are used. A tradeoff in validity evidence occurs when an examining board uses a test of less fidelity to achieve greater sampling of the intended domain of criterion behaviors. It is difficult to have high fidelity and adequate sampling at the same time. Thus, the test sponsor/developer often compromises to achieve high fidelity with less effective sampling or lower fidelity and has more effective sampling. WREB has worked out a reasonable compromise between these two opposing tensions in test design involving content. More information about these tests is found in the *Candidate Guide* (WREB 2009). WREB has released two statements discussing the pros and cons of using manikins instead of actual patients (http://www.wreb.org/Files/Articles/pos-supt.pdf and http://www.wreb.org/Files/Articles/CrownPrepPosition2009.pdf. WREB's conclusion is to maintain a high fidelity testing situation. A validity argument has yet to be convincing that a manikin will provide a more valid test score interpretation. The evaluation of the reasonableness of such compromises is best left to a panel of SMEs who have the appropriate professional judgment to make such an evaluation

Conclusion

WREB has assembled a comprehensive and appropriate body of evidence supporting the content of the five tests comprising the *Dental Examination*. As the tests in the current examination includes parts of the previous compensatory-scored examination, there is an historical continuity in the content of the tests that suggest stability with respect to content.

Examiner Training

WREB has documentation showing how examiners are prepared for the examination. Each examiner receives an updated *Examiner Manual* (WREB, 2009c). This manual contains general and specific information. The general information addresses basic issues that apply to all clinical tests. The *Examiner Manual* has sections for each clinically-based test. Additional on-site calibration training and testing is accomplished on the day prior to the beginning of each test administration. The specific training information for each of the relevant tests and sub-test is provided in subsequent sections.

Conclusion

WREB has documented evidence supporting its extensive examiner training activities. This system of training has evolved over many years, and appropriate committees serve to improve it continuously.

Examination Administration

As the *Dental Examination* was revised for 2009 administrations, evidence was presented about how dental schools and their deans were informed of this fact (Ramos, October 30, 2008; October 14, 2009a; October 14, 2009b; October 14, 2009c). The most comprehensive and authoritative source of information about examination administration is the *Policy and Procedures Manual* (WREB, 2009c). This volume describes the application process, how examination results are prepared, travel and shipping, and administrative issues and procedures.

Overall planning for administration is documented showing how test sites and its coordinators need to prepare for examination (WREB, October 14, 2009a, October 14, 2009b, October 14, 2009c). Every candidate receives a letter informing the candidate about the orientation day and the three clinical testing days and, also, the computer simulation test and subtest.

Conclusion

As with examiner training, WREB's extensive and the use of appropriate SME committees has served continuously to improve test administration. Considerable documentation is available to support the conclusion that examination administration is among the strengths of this examination program.

Comparability

Comparability in this examination is considered test by test. Clinical performance tests have standardized administration and scoring using a standardized scale with its cut score set at three on a five-point scale. The *Operative, Endodontics*, one part of the *Periodontal* subtest is scored by trained, validated examiners. Thus, scores from the first two tests and the subtest can be considered as having comparable score scales from test site to test site across years due to the standardization that is evident.

The *Patient Assessment and Treatment Planning* test is administered by computer, but the scoring is done by trained, validated examiners. Thus, this test score scale is also standardized. Comparability of the scale is based on an argument, not on an equating procedure. That argument is that all aspects of the test are standardized, so the score scale is constant from test site to test site across years. Equating methods used for multiple-choice tests are unsuitable for standardized performance tests.

The *Prosthodontics* and the second part of the *Periodontal* test (*Periodontal assessment/diagnosis*) are presented in a multiple-choice format and are objectively scored. These two tests are subject to test score equating to ensure that the cut score of 75 on a 100-point scale is standardized (equal) for all candidates. More information is provided in the parts of this report that deal with unique tests. Briefly, an equipercentile equating is done based on the assumption that different test forms are randomly administered to comparable samples. Data is presented later that shows that performance across content-equivalent test forms has comparable difficulty, so there is evidence of equivalence of these test forms and adequate sampling.

Conclusion

The two tests and one subtest that use examiner judgment on rating scales appear to have satisfied standards for comparability so that test scores reported from site to site and from one year to the next can be interpreted equivalently. That is, the test score scale is constant.

For two other tests and the second part of the *Periodontal* test, comparability will be discussed in appropriate forthcoming sections of this evaluation report. In this subsequent analysis of comparability, data will be presented that effectively shows that the test forms appear to be on the same test score scale.

Standard Setting

WREB determined that in its clinical tests and the *Periodontal* clinical sub-test, a five-point rating scale would be used (Ramos, June 23, 2009). A consensus decision of the Examination Review Committee was that the cut score would be three on this scale. This decision was based on Standard 14.17 in the *Standards for Educational and Psychological Testing* (AERA, et al, 1999), which states "The cut score should be determined by careful analysis and judgment of acceptable performance." Consequently, rather than using a percentage for the cut score (passing level), WREB subcommittees have developed scoring rubric by which a rating of three is passing. As three examiners rate performance, the median score is used instead of the mean. As a measure of central tendency, the mean is often preferred if the sample is large and the distribution is normal. As the sample size is small and the distribution is skewed, the median is a more representative measure of central tendency than the mean. Thus, the median of the three ratings is used when determining the test score. It should be noted that penalty scores have also been revised to align with the new score reporting. The point assignments, for each scored procedure, are published in the *Candidate Guide* (WREB, 2009b). A score of three is a minimum passing level, whereas a score of four or five represents better performance.

The *Periodontal Assessment/Diagnosis* sub-test and the *Prosthodontics* test are computer-scored and have 100 points for each of the sections. The passing standard is 0.75. Because the different forms of the computerized sections are equated, a score of 0.75 represents the same passing level of performance for each form. These passing standards are consistent with Standard 14.17. The passing score for the Periodontal test is 0.75 on a scale from 0.00 to 1.00, where the results of the test subtests are combined (see Hammond, February 2010).

Conclusion

The standard of three on a five-point rating scale is a function of the descriptive rating scales used. This is a consensus-based judgmental standard that is defensible. Thus, three tests and one subtest have met this standard. These are (1) *Operative*, (2) *Endodontics*, (3) the *Patient Assessment and Treatment Planning* test.

Reporting

The *Candidate Guide* (WREB, 2009b) provides description about scoring. The score report is designed to reveal candidate performance in all aspects of the examination on a point basis against possible points to be earned. Confidentiality of candidates' results is ensured. Candidates graduating from dental schools have the option of withholding their score report from their school. No other parties have access to these scores, unless expressly designated by the candidate. WREB contractually provides reports to member states. Schools are sent reports unless students do not wish to have the schools receive their scores. Standards 5.13, 8.5, and 11.14 from the *Standards* (AERA, *et al.*, 1999) are clear about this need for useful information to be reported and confidentiality**.**

A candidate score report should present test results clearly and effectively. The score report should help candidates understand the scoring procedure and the meaning of scores on the report that comprise the total score. Score reports are confidential and are not public documents. An inspection of a typical school score report and the individual score report shows clear comprehensive candidate performance that includes previous attempts, point deductions, and performance points for each test. For passing candidates, WREB simply announces whether the candidate has passed or failed. For failing candidates, WREB provides a one-page detailed diagnostic score report for all five tests. The *Operative* test has the greatest detail, and the *Prosthodontics* test has the simplest report, a percentage score.

A dental school report lists the names of the candidates and their scores on each of the five tests and the pass/fail decision reached. The first three tests are presented on the five-point scale, where three is passing. The *Periodontal Assessment/Diagnosis* subtest is reported in percentage where 75 is the cut score. The *Prosthodontics* test is also presented in the percentage correct scale where 75 is the cut score.

Conclusion

WREB's score reports are clear and useful to candidates because diagnostic information is provided on strengths and weaknesses. Candidates' rights regarding confidentiality are respected. WREB fully meets all standards bearing on score reports. No recommendations are offered for improvement.

Candidate and Patient Guide and Rights

The *Candidate Guide* (WREB, 2009b) has been cited often in this report. This booklet contains essential information for candidates. The table of contents for this 108-page booklet provides general information, performance evaluation information, patient criteria, and examination procedures. The booklet is published each year and is updated as needed. Besides this guide, WREB's webpage is helpful, and if candidates prefer can contact the WREB office by phone or by email.

WREB also issues regular newsletters of general interest to dental students (WREB, Summer, 2008; Fall, 2008; Winter/Spring, 2009; Fall, 2009a; Fall 2009b). A review of these newsletters will reveal that candidates are informed about various issues they might encounter in their attempt to pass this important examination. Some of these issues are appeals process, score information, characteristics of successful candidates, advice on test preparation, choosing patients, application process, scoring procedures, and examination calendars.

If candidates have special needs as provided in the Americans with Disabilities Act, WREB provides reasonable and appropriate accommodations (see *Policy and Procedures Manual*, WREB, 2009c).

Patients are part of the examination process. Considerable attention is given to patient rights and care. An annual survey is conducted of patients inquiring about their treatment and satisfaction (http://www.wreb.org/Files/Articles/PtSurveyart.pdf)

Conclusion

Information and references to information about candidates' rights have been presented here and appear in the archive. The *Standards* (AERA, *et al.*, 1999) are clear in chapter 8 about the rights and responsibilities of test takers. Standards 8.1 and 8.2 address keeping candidates informed about the test. WREB meets these standards fully. WREB is commended for its *Candidate Guide* (WREB, 2009) and its frequent and informative newsletters. These documents are exemplary as communication tools for candidates, and these documents also provides a variety of well-documented validity evidence that assures the candidates and others about the quality of this testing program.

Security

The *WREB Policy and Procedures Manual* (WREB 2009c) discusses security. WREB has security processes and policies for both technology hardware and software. Organization data is stored and processed on servers, which run from locked rooms. The server rooms are secured using keypad entry locks, limited to executive and information technology team access. The WREB office suite is locked after normal business hours and only accessible after hours with key card access. Key cards are monitored by building security system. Data regarding office access and video surveillance of building entry ways is monitored and saved by building management company. Besides server security, electronic scoring system hardware is also stored in locked limited keypad access rooms.

As far as organization data is concerned, because data is stored and processed from central servers, critical files are not stored on individual PCs. A data backup process runs several times per week locally, and once per week offsite. Access in and out of the WREB internal network is guarded by hardware and software firewalls. In case of travel or emergency, WREB staff may have access to office data files remotely. However, access is restricted to specific user roles, only available as needed and facilitated by WREB information technology team.

Offsite critical data is also copied for redundancy and secure. The WREB website is hosted offsite. Candidate data that is collected through the website is encrypted and verified with licensed SSL certificate. Credit card information from online candidate registrations is not available to WREB staff or saved in a database. Candidate-specific information is available on the website using candidates' individual login accounts. A secured section of the website is also available for examiners who have been approved for access by WREB staff after verifying their access rights to the information.

With any high-stakes examination, the temptation to cheat is great. WREB requires each candidate to be clearly and accurately identified and monitored during the examination process. Each candidate must have their patient qualified. Failure to do this well or do it at all puts candidates in jeopardy of failing.

It is extremely unlikely that a single examiner could undermine the validity of any examination. Self-interest or other factors may contribute to unwarranted ratings. Although this kind of behavior is unlikely to be considered cheating, it is undesirable and a threat to validity. This problem is unlikely in the WREB environment where examining team assignments are carefully trained and monitored during the administration to reduce this possibility. The integrity of examiners is discussed in the *Dental Examiners Manual* (WREB 2009d). Conflicts of interest with candidates are monitored, and examiners are asked to recuse themselves if potential conflicts exist. Finally, as reported previously, WREB has high standards for selecting examiners (WREB, 2009a).

Conclusion

The procedures for security established over many years are well documented (WREB, 2009c). WREB has provided excellent validity evidence bearing on security. It meets the standard 5.9 (*Standards*, AERA, *et al*., 1999).

PART VII: OPERATIVE VALIDITY EVIDENCE

Each candidate completes two restorative procedures chosen from four options: (1) direct posterior Class II amalgam restoration, (2) direct posterior Class II composite restoration, (3) direct anterior class III composite restoration, and (4) indirect posterior class II cast gold restoration. For each restorative procedure, the candidate does a preparation and a finish. For each preparation and finish, three examiners provide three ratings of performance.

Test Development for the 2009 Test

The transition of the *Operative* test to conjunctive scoring is well documented in committee minutes (May 19-20, 2006; September 8-9, 2006; November 30-December 1, 2007; February 2-3, 2007; September 7-9, 2007; February 29-March 1, 2008; May 9-10, 2008; September 5-6, 2008; March 6, 2009; May 8-9, 2006; September 11-12, 2009). The *Operative* test is continually being evaluated, and these minutes reveal many fine-tuning adjustments to the examination with the intent of improving this test over these years.

Item Quality

The tasks to be performance and the descriptive rating scales are presented in the *Candidate Guide* (WREB, 2009b, pp. 30-47). As noted in the content-related validity section, the tasks are those done by dentists in practice, so each task has high fidelity with the criterion behavior it is supposed to represent. The five-point descriptive rating scales are well written and have the added feature of having the median rating (three) represent the cut score for making pass/fail decision. Thus, each examiner determines if the minimum passing has been met and then assigns a rating of three, four, or five according to the perceived level of performance exhibited by the candidate.

Training of Scorers and Scoring

The method of scoring for the 2009 *Dental Examination* was presented to the Examination Review Committee on July 8, 2008. This document provides information about the complex scoring arrived by the examination committee. The training materials for *Operative* test are extensive entailing nine documents that address a tutorial overview, a self-test, a self-test key, a preparation tutorial and a preparation self-test, and a preparation self-test key, and an operative finish tutorial, self-test, and key. The *Examiner Manual* (WREB 2009c, pp. 33-72) provides extensive information about what examiners need to know and do for the operative test.

Scoring is very complex. Candidates must choose two among four procedures, and in the sample of data provided, no candidate chose an indirect posterior class II (Cast Gold). The main scoring is based on five-point rating scales, but deductions play a significant role in scoring. Although the bottom of the rating scale is one, it is possible for candidates to have zero scores due to deductions. As all deductions are validated by consensus and a scoring rule where the median is the value used for a decision, the resulting score is considered in rater consistency and reliability analyses in this evaluation.

Is Their Bias in Choice?

As candidates can choose among four procedures (amalgam, composite, gold, and class III composite), is bias introduced when each candidate chooses? The means and standard deviations for preparation and finish are presented in Table 4. A two-way analysis of variance with repeated measures was used to analyze the data. As this analysis is very powerful and the sample size is very large, statistically significant results are very likely even if differences are very small. As a safeguard against drawing a false conclusion, all results were subject to an evaluation of effect size (standardized mean difference) to decide if any differences are consequential.

A main effect was observed for type of tooth chosen ($F= 40.3$, $df = 2, 4467$, $p < 0.0001$). However, this result accounted for only 1.7% of the variance. A difference was also found for the difference between preparation and finish ($F = 6.39$; $df = 1, 4467$; $p = 0.011$. However, this difference has a very small effect size and is not relevant to the question asked. A significant interaction was detected for the type of tooth and preparation/finish ($F = 58.18, 2, 4467$, $p < 0.0001$). Looking at the two tables, the finish portion of the test for the class 3 composite tooth has higher scores than all other scores. The mean difference between the finish mean for composite 3 and composite finish is 0.34, which is substantial.

Table 4: Descriptive Statistics for Three Choices

Preparation	Sample	Minimum	Maximum	Mean	S. D.
Amalgam	1,750	1.15	5.00	3.62	0.676
Composite	1,787	0.60	5.00	3.69	0.697
Class 3 Composite	943	0.75	5.00	3.69	0.698

Finish	Sample	Minimum	Maximum	Mean	S. D.
Amalgam	1,745	1.27	5.00	3.57	0.63
Composite	1,785	1.36	5.00	3.50	0.59
Class 3 Composite	940	1.27	5.00	3.84	0.69

As Table 5 shows, reliability estimates are very high and higher for finish when compared with preparation.

Table 5: Reliability Estimates for Preparation and Finish for Three Procedures

	Preparation	Finish
Amalgam	0.907	0.976
Composite	0.907	0.974
Class 3 Composite	0.926	0.987

Examiner Consistency and Reliability

Examiner consistency and reliability are functionally related. If examiner consistency is high, reliability will be high if the dimensionality of the test is singular. If it is not singular, than other methods exist for estimating reliability, but we would also expect higher reliability for subscores on the test.

Examiner Consistency

Table 6 presents an index of rater consistency for each rating in the preparation and finish subtests. These indexes of internal consistency among raters (using coefficient alpha) are moderate. For five-point, descriptive rating scales, these indexes are sufficiently high to support reliable preparation, finish and total test scores. The proof of this assertion is found in the next section.

Table 6: Index of Rater Consistency (including deductions)

Preparation–First Tooth			Finish–First Tooth		
Subscale	Weight	Index	Subscale	Weight	Index
outline/extension	46%	0.762	anatomical form	36.5%	0.743
internal form	39%	0.784	margins	36.5%	0.705
operative environment	15%	0.830	finish/function/damage	27.0%	0.706

Preparation–Second Tooth			Finish–Second Tooth		
Subscale	Weight	Index	Subscale	Weight	Index
outline/extension	46%	0.772	anatomical form	36.5%	0.782
internal form	39%	0.793	margins	36.5%	0.706
operative environment	15%	0.814	finish/function/damage	27.0%	0.705

Reliability Estimate of Operative and Finish Scores

The reliability of the preparation and finish subscores on the first tooth were 0.858 and 0.875 respectively. The reliability of preparation and finish subscores on the second tooth were 0.861 and 0.888 respectively. These coefficients are very high for clinical performance-based tests using descriptive rating scales.

Total Operative Reliability

As preparation and finish are different, coefficient alpha would not be appropriate. A stratified alpha is preferred when subscores of a test are distinctly different, which it is in this instance (see Haertel, 2006). The stratified alpha coefficient was 0.924. The standard error of measurement is 0.120. As reliability is high, random measurement error is very small. The cut point is three, so we establish a confidence interval around the cut point of 0.120. The range of uncertainty is 2.88 to 3.12. A total of 274 candidates had scores in this area of uncertainty, which is about 12%. In all tests with a cut score, this range of uncertainty exists and there is no way to ascertain the true status of candidates. An observed score of 3.12 might have a true score below 3.00, and an observed score of 2.88 might have a true score above 3.00. The strategies for dealing with this eventuality are to (1) retest candidates because they are too close to the borderline to decide with great confidence, (2) provide a standard error of "grace" and allow anyone with 2.88 or higher to pass, or (3) provide a standard error of "regret" and assign anyone below 3.12 to fail so that there is no chance that someone who is less competent will pass. Another strategy is to redesign the test to increase observations and thereby increase reliability, but as reliability is very high, it would be difficult to increase reliability and make this zone of uncertainty much smaller.

If item response theory scaling were done, the conditional standard error would be estimated and this is likely to be smaller and the uncertainty band would be smaller. But this is a theoretical issue with little bearing on the problem of deciding who will pass or fail.

Conclusion

The operative examination is a very complex one. The content of this test involved high fidelity performance based on the results of the practice analysis and interpreted by a committee of SMEs. The test was converted into an independent, conjunctive pass/fail test. The candidate chooses two types of operative procedures from a list of four. Apparently gold was not chosen in this sample of candidates. The procedures for this test are very clearly described in the *Candidate Guide* (WREB, 2009b). Performance tasks were weighted by the SMEs appropriately via consensus. The descriptive rating scales were developed to enable ratings. Training appears sufficient. The scoring criteria are also very clearly stated. Scoring appears very consistent, and sub score and total score reliability is very high. Standard error is very small for the total score. Overall, this test deserves high marks for validity.

One potential source of CIV is the Class 3 Composite test. Candidates who choose this technique scored higher on the finish part of the test. Possible reasons for this offer only speculation. The committee should study this problem. A validity study might provide more insight into whether candidates' score are biased or the byproduct of differences in training at various dental schools. Another approach is to standardize the test and limit candidates to only two techniques.

PART VIII: PERIODONTAL VALIDITY EVIDENCE

The Periodontal test has two subtests: the Periodontal Treatment and the Periodontal Assessment/Diagnosis. Because these two subtests are developed independently, each will be evaluated in its own section.

The practice analysis conducted by WREB (1999) and again in 2007 (WREB, May 15, 2007) provided the content basis for each subtest. These subtests have judged by SMEs to be complementary in covering the knowledge, skills, and abilities intended for Periodontal test. Minutes for periodic meetings provide documentation for the content decisions and the improvement of the Periodontal Treatment subtest and the design and development of the Periodontal/Diagnosis subtest (Baker, March 14, 2003; June 24, 2005, CRDTS/SRTA/WREB, November 14-15, 2003, February 14-15, 2003, March 27-28, 2004, July 10-11, 2004, October 29, 2004, January 28-29, 2005, June 24-25, 2005; WREB, February 1-2, 2003; CSW, September 9-10, 2005; December 10, 2005; Hammond, January 17, 2007; WREB April 9, 2009).

Periodontal Treatment Subtest

The *Periodontal Treatment* subtest requires the candidate to perform scaling and root planing on one or two quadrants of the mouth. The *Dental Candidate Guide* (WREB, 2009b, pp. 65-73) describes in detail the specific procedures and scoring of the subtest. This subtest has been used for many years and has undergone minor revisions throughout its history. A subcommittee exists for developing and improving the periodontal treatment test. This test is perceived to be high-fidelity due to the performance nature of the tasks with an actual patient.

Item Quality

The instructions are very detailed and clearly presented. Scoring is described on page 70 of the candidate guide. Three reasons are listed for point deductions. Each point deduction is validated by the rule that all point deductions are validated by two examiners. The scoring is based upon the number of calculus pieces remaining after the procedure is completed. A candidate's score can range from 25% to 100% based on the amount of calculus remaining. Items appear to have high fidelity with the criterion behavior that the sub-test is supposed to represent. Meeting minutes previously cited provide many details about sub-tests development and the development of items.

Rater Consistency and Reliability

Three examiners score eight tooth surfaces. For this data set, 6,314 pair comparisons, and the percentage lack of agreement for each area is presented below. Rating patterns are as follows: 000, 111 (complete agreement) 001, 010, 011, 100, 101, 110 (partial agreement). For complete agreement we have three agreements. For partial agreement (validation), we have two of three agreements. Thus, this agreement scale varies between 66.7% and 100%

Surface 1	Surface 2	Surface 3	Surface 4	Surface 5	Surface 6	Surface 7	Surface 8
94.3%	94.4%	93.9%	95.3%	95.2%	95.4%	96.2%	96.9%

The level of rater consistency is very high. The obtained score is a consensus. Where this is 100% agreement, the total score (1 or 0) is affirmed. Where there is two of three, the median value is used.

The reliability coefficient is 0.574. The mean of these scores is 91.8% and the standard deviation is 0.104. The standard error of measurement is 0.067. An inspection of the highly skewed distribution shows that very few candidates have calculus remaining that justifies a low score on this sub-section of the test. A conclusion is that reliability is attenuated due to the highly skewed distribution of scores and the considerable restriction of range. Because the standard error of measurement is so small, precision of these scores is very high.

No conditional standard error was estimated. As discussed previously, this statistic is desirable but is not easily obtainable under current conditions. If the test were scaled using item response theory, conditional standard errors would be possible. Such a statistic would have no impact on the conclusion drawn that the *Periodontal* test scores are precise.

Training of Scorers and Scoring

The *Examiners Manual* (WREB, 2009c, pp. 89-100) provides general information, treatment criteria, check-in criteria, reasons for patient rejection, and scoring information. The procedure for the training and scoring appears to be very sound.

Comparability

As noted previously, the *Periodontal Treatment* subtest is standardized in administration, training of examiners, and scoring. Therefore the scale used for interpreting candidate performance is uniform across candidates within any one administration and across all administrations. That is, evidence of comparability is sufficient.

Conclusion

This sub-test is very well developed and provides validly interpretable results. The restriction in the range of performance and the high uniform performance results in a low reliability estimate, but the standard error of measurement is small.

Periodontal Assessment/Diagnosis Subtest

This subtest measures "knowledge and judgments" dealing with periodontal care. The *Dental Candidate Guide* (WREB, 2009b, p. 4) gives a brief description of this subtest. This subtest, developed by CSW Computer simulations, is administered and objectively scored by computer.

This subtest has been under development for some time (Baker, November 11, 2004; CRDTS/SRTA/WREB, November 14-15, 2003, February 14-15, 2003, March 27-28, 2004, July 10-11, 2004; CSW, December 10, 2005; WREB, February 1-2, 2003). More information about this subtest development can be found in the technical report by Hammond and Buckendahl (March 2007, June 2008), Hammond (June 2009), and more recently in a report by WREB (February 2010). These technical reports provide validity evidence.

Item Quality

A history of the item development and validation can be found in technical reports (Hammond and Buckendahl, March 2007, June 2008; Hammond, June 2009; February 2010). Patient cases were used as the basis for developing items. These cases included medical history, oral history, assessment, prognosis, treatment plan, and reevaluation. Items were field tested with dental students. A review of a large sample of items shows that standard item-writing guidelines were followed.

An inspection of item analysis results revealed that items have high p-values reflecting the high level of performance of candidates. Because of the negative skewness of subtest scores, item discrimination indexes are attenuated. This is not a reflection of item quality but a reflection of high performance by these candidates. A discrimination index works best when high-scoring and low-scoring candidates take the test. In this test, very few candidates scored low.

Test Administration

The test is administered by computer by the Pearson Vue at any of their testing centers. The *Candidate Guide* (WREB, 2009b) provides information about scheduling the subtest and its administration. Each candidate is randomly assigned a clinical patient and patient file.

Scaling and Comparability

As this subtest consists of 50 items, the test score scale should be equivalent for each test form. CSW uses equipercentile equating to achieve equivalent scale for each test form. As the test forms are randomly assigned, an assumption of equivalent content and difficulty is claimed. The results shown below in Table 7 support that claim. Very small score adjustments are made in the scores using the equipercentile method.

Table 7
Descriptive Statistics for the Periodontal Tests for 2009

Form	Sample	Mean	S. D.	Nitems	Discrim	Alpha	Alpha x2
pe-39an	339	0.846	0.090	39	0.191	0.658	0.794
pe-39ag	303	0.859	0.084	39	0.168	0.635	0.777
pe-42	624	0.865	0.081	42	0.162	0.635	0.777
peaj-42	312	0.868	0.081	l42	0.167	0.644	0.783
peal-42	312	0.862	0.092	42	0.158	0.626	0.769
pe-44-3	640	0.874	0.78	44	0.156	0.639	0.780
pe44-3be	298	0.872	0.077	44	0.146	0.614	0.761
pe44-3bi	342	0.875	0.080	44	0.164	0.659	0.794
pe46-4	603	0.881	0.072	46	0.136	0.597	0.747
pe46-4bi	311	0.874	0.075	46	0.145	0.616	0.762
pe46-4bo	292	0.888	0.067	46	0.122	0.566	0.723

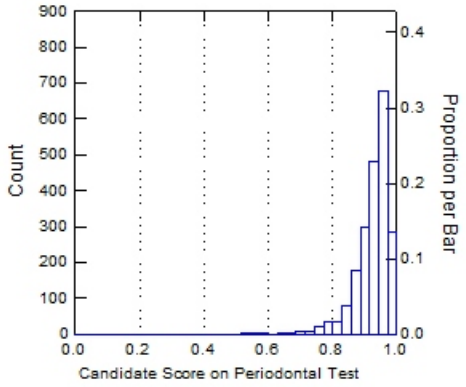
Reliability

Hammond and Buckendahl (March 2007; June 2008) and Hammond (June 2009) reported that alpha reliability estimates were typically very low. Mean scores ranged ranged from 80% to almost 90%. Standard deviations showed very little variation among scores. The decision consistency indexes were very high. As with item discrimination, the extent of low reliability coefficients is not an instance of low reliability but more of a ceiling imposed by a lack of variability in the performance of candidates, who appear to be similar in ability. The last column shows what the reliability would be if the test were doubled in length. High-stakes sub-tests like this one should have higher reliability. Another factor that negatively affects reliability estimates is low item discrimination values. Using items with p-values of 0.90 or higher tends to lower item discrimination. In other words, lower discrimination lowers reliability. Because the content of these items is critical to the skills being measured, these items must be used. Low reliability is not necessarily a threat to validity. The standard error of measurement is also small. So the risk of misclassifying a candidate whose score is at or near the cut score is very small.

Periodontal Total Score Reliability

As these two subtest are equally weighted and combined to form a total score, total score reliability can be estimated using coefficient alpha. A scatter plot of the scores from the two

subtests show virtually no relationship. As the content in the two subtests is complementary and performance is very high, reliability cannot be estimated due to the considerable restriction in range and skewness. Relying on the most conservative estimate involving coefficient alpha, the standard error is 0.002. The cut score is 0.75. As shown in the histogram below, very few candidates score this low on the test. Because of this unusual distribution, pass/fail decisions are extremely reliable.



Conclusions

1. The *Periodontal Treatment* sub-test measures the effectiveness of calculus removal. Candidates perform very well on this sub-test. As a result, there is little variation in scores, and reliability estimate is attenuated. Nonetheless, there is considerable evidence from very high rater consistency that there is much consistency in these judgments.

2. For the *Periodontal Assessment/Diagnosis* sub-test, validity evidence seems adequate but documentation can be improved. The test length should be standardized (e.g., 50 items or 75 items). The longer test length is preferred due to a higher expected test score reliability.

   Overall, the Periodontal test appears to have useful validity evidence.

PART IX: ENDODONTICS VALIDITY EVIDENCE

The Endodontics test provides for the treatment of two extracted teeth: one anterior and one multi-canal posterior. For each tooth, access is evaluated and is weighted 37.5%, and condensation is evaluated and is weighted 62.5%. Descriptive rating scales appear on page 62 of the *Dental Candidate Guide* (WREB, 2009b). If a tooth is rejected for examination on the first or second trial, there is no point deduction. However, on the third rejection, no additional submissions are allowed and no points are awarded. Criteria for rejection are provided on page 60 of the candidate guide.

Test Development for the 2009 Endodontics Test

The Endodontic Committee meets regularly to discuss issues and policy related to the new 2009 test (e.g., July 19, 2006; May 9-10, 2008; August, 22-23, 2008; May 15-16, 2009). Included in these minutes are discussions and a consensus reached on scoring criteria and administration procedures ensuring high reliability for the new conjunctive scoring system.

Item Quality

The procedures for administration and scoring are provided in the *Candidate Guide* (WREB, 2009b, pp. 48-69). This section provides a test overview, which teeth to select, which kinds of teeth are not acceptable, how to prepare teeth, radiograph submission, treatment, and after treatment. The test is well supported by previous test administration procedures with the major difference being that the test score for endodontics is now one of the conjunctive scores used for a pass or fail decision.

Reliability and Examiner Consistency

As each of the three examiners does four evaluations, a useful index of agreement is coefficient alpha. For the four evaluations, these coefficients were 0.813, 0.841, 0.841, and 0.847. These coefficients are very good considering the use of a five-point, descriptive rating scale and only three observations per scale.

For the four subtests of the *Endodontics* test (access and condensation for each of the two teeth), the alpha estimate of reliability was 0.887. This result is very high for a clinical performance test. The standard error of measurement for this 0.430. The confidence interval (band of uncertainty) ranges from 2.570 to 3.430. Those candidates whose ratings fall between these two values are uncertain regarding their true pass/fail status.

Training of Scorers and Scoring

The training materials for these examiners include an orientation and calibration session and another session devoted to basic principles of endodontic access followed by a self-test and a self-test key. The *Examiner Manual* (WREB, 2009c, pp. 73-88) provides general information, approval criteria and procedures, scoring, definitions, reference material, and a floor examiner job description.

A total score is the median and not the mean of three examiner ratings. This procedure is defensible because for small samples, the median is the most representative measure of central tendency.

Conclusion

The Endodontics test is well validated. Reliability is high. Examiner consistency is high. No recommendations are offered for this test.

PART X: PROSTHODONTICS VALIDITY EVIDENCE

This test consists of 50 randomly selected items from an item bank. For each test there is a three dimensional model. The 2010 test specifications call for nine complete dentures items, 18 removable dentures items, 16 fixed items, and seven implant items. The *Dental Candidate Guide* (WREB, 2009b, pp. 102-103) provides more information about this examination. Technical reports provide information about this test (Hammond & Buckendahl, March 2007; Maarch 2008; Hammond, June 2009; WREB, February 2010).

Test Development for 2009

WREB's development of this test has a well-documented history up to 2005 (e. g., August 15-16, 1998; January 9-10, 1999; April 17, 1999, April 23, 1999; September 25-26, 1999; December 11-12, 1999; August 18-19, 2001; November 10-11, 2001; February 15, 2002; August 2-3, 2002; October 19, 2002; February 14, 2003; July 25-27, 2003). At the initial meeting, test specifications and the passing score were discussed. The need for high reliability was emphasized and a report of recent progress was made. At one of these meetings it was decided to use computer simulations (August 2-3, 2002).

The test specifications were established at one of these meetings (WREB, September 25-26, 1999) and later revised (CSW, October 27, 2005; WREB, August 18-19, 2001). The periodic practice analysis is used to review and update test specifications. At the many meetings of this committee, all aspects of test development were discussed and decisions made that established the basis for the current test. According to Hammond and Buckendahl (2007), the practice analysis led to the formation of current test specifications. The latest report updates the development of this test for the conjunctive scoring (WREB, February 2010).

Item Quality

Items were developed by CSW subject-matter committees consisting of dental school faculty members and practicing dentists. Each test item was developed to measure specific knowledge and skills drawing from the patient problem and other information presented. A sample item is provided in the *Dental Candidate Guide* (p. 102). Technical reports provide background on item development (Hammond and Buckendahl, March 2007; March 2008; Hammond, June 2009; WREB, February, 2010). More detail is suggested in future reports on item quality.

Results of test and item analyses are shown in the Table 8. The means of the test are similar, which provides some evidence that the test has equivalent difficulty. Because the tests conform to the test specifications, the content is equivalent from form to form.

Table 8
Descriptive Statistics for the Prostodontics Tests for 2009

Form	Mean	S. D.	Nitems	Discrim	Alpha	Alpha x 2
pr-28	0.819	0.098	28	0.154	0.525	0.688
pr-31	0.831	0.086	31	0.113	0.472	0.641
pr-33	0.824	0.080	33	0.099	0.429	0.600
pr-38	0.804	0.085	38	0.116	0.488	0.656

Scoring

The test is objectively scored. Non-performing items were removed before scoring candidates as part of key validation.

Standard Setting

The Ebel method for standard setting was used to set the cut score this examination (Hammond & Buckendahl, 2007). This method also involves an important step in item development and validating, a rating of each item's relevance.

Scaling for Comparability

Each candidate test is randomly parallel, so test difficulty should be constant from form to form. As shown in Table 8, the means and standard deviations are approximately equivalent. Adjustments due to equipercentile equating are very small, which is very desirable. Thus, there is good evidence supporting comparability of test forms and uniformity of the test score scale for making pass/fail decisions.

Reliability

In technical reports for this test, Hammond and Buckendahl (March 2007; March 2008; Hammond, 2009; WREB, February 2010) very low reliability coefficients were reported. Table 8 provides reliability estimate results for 2009. One reason for low coefficients is a restriction in the range of scores. The test results show a high level of performance. The standard error for one test form is 0.055, which is very small. Few candidates' scores fall at or near the cut score of 0.75, so the risk of being misclassified due to random errors of measurement is very small.

Another cause of low reliability might be the number of items. In Table 8, estimated reliabilities for doubling the length of the test were presented using the Spearman-Brown formula. As shown in that table, even with longer test lengths, reliability does not reach a desirable level. The test would need to be both longer and have higher discriminating items to improve reliability estimates.

Conclusion

The Prostodontics test has received considerable attention from 1999 to 2005, at which point documentation of meetings is sparse. Technical reports by Hammond and Buckendahl (March 2007; March 2008; Hammond June 2009; WREB, February 2010) provide validity evidence.

Some recommendations for improvement are offered:

1. The test should have a standardized length. A longer test is recommended to achieve higher reliability.

2. Documentation of validity evidence should be improved. Meeting minutes should be completed and dated to assure WREB that the test is being improved. In turn, WREB can assure its participating states that the information provided in this test is validly interpreted and used.

The test appears to provide validly interpretable scores.

PART XI: PATIENT ASSESSMENT AND TREATMENT PLANNING (PATP) VALIDITY EVIDENCE

This test is a computer simulation using patient case material. The candidate has one hour to assess images and consider patient information. The candidate must then complete and submit a treatment plan. Each candidate is randomly assigned a patient. Examiners score results.

Test Development for 2009

Meeting minutes show the systematic development of the PATP conjunctively-scored test (October 28-29, 2005; March 3-4, 2006; July 7-8, 2006; August 25-26, 2006; December 1-2, 2006; February 9-10, 2007; June 1-2, 2007; December 7, 2007; May 30-31, 2008; August 22-23, 2008; January 16-17, 2009; June 19-20, 2009). These documents provide detail about the maintenance of the current examination and the development of the new conjunctively-scored test, which is based on the previous validated examination. The *Dental Candidate Guide* (WREB, 2009) provides information about this test.

Item Quality

Candidates receive a computer file containing eight pieces of information about a patient. The candidate must prepare a treatment plan. Criteria for the treatment plan are presented in the *Candidate Guide* (WREB, 2009). The scoring guide is presented on page 81 of the *Candidate Guide*. Descriptive rating scales are presented on pages 82 and 83. The tasks and rating scales are well developed. The tasks are based on the results of recent and past practice analyses.

Examiner Consistency and Reliability

For each of the 11 criteria being evaluated, three examiners rated each using a five-point, descriptive rating scale. An index of rater consistency is reported in the table below.

Key 1	Key 2	Key 3	Key 4	Key 5	Mod	Inc	NonEx	Seq	Clear	Main
0.944	0.952	0.967	0.959	0.946	0.901	0.795	0.751	0.532	0.626	0.928

The inter-examiner agreement coefficients show some variability. These indexes look at total examiner agreement. As WREB uses a scoring rule where the median instead of the mean is used, the above coefficients may not represent the facts of examiner consistency perfectly. Nonetheless, the above coefficients are accurate in showing the high degree of overall agreement among the examiners.

Using coefficient alpha, the reliability estimate for the 11 scores comprising a total score is 0.955. The standard error for the total score is 0.102, which is very small. A conditional standard error at the cut score (75) would be preferable, but one would have to use another item response theory and a more complicated method of scaling. The conditional standard error would be slightly smaller most likely. However, this fact is inconsequential given reliability is so high and the standard error is so small.

Training of Scorers and Scoring

The training materials for examiners include a PATP tutorial and a self-test and self-test key. Three examiners independently score the treatment plan. The *Examiner Manual* (WREB, 2009, pp. 101-112) provides several sections providing information about criteria and how to score with many examples. Examiners receive on-site calibration training and are tested before each scoring session. Examiner consistency might be improved with more effective training for some scales of this test, but considering the overall reliability, this improvement would have very little overall effect on the validity of test scores.

The *Dental Candidate Guide* (WREB, 2009b) provides a very detailed description of the procedures and scoring criteria for this test. In the minutes of the PATP Planning Committee (July 8-9, 2005), the move from a compensatory to a conjunctive scoring system was underway.

Conclusion

The validity of test score interpretation and use for the PATP test is excellent. This test can be used confidently for making pass/fail decisions as needed. No recommendations are offered for this test's improvement.

VII: Summative Evaluation and Recommendations

The Dental Examination consists of five independent tests, each of which is subject to validation. This final section summarizes findings and recommendations of this evaluation.

Operative Test

As noted previously, this test is very complex. Because of that fact, it is very challenging to score reliably. This test has excellent qualities and a very supportive body of validity evidence. This test is very valid for making pass/fail decisions for the content that this test is supposed to measure. No recommendations were offered for its improvement.

Periodontal Test

This test consists of two subtests that measure very different content. As a result, the two subtests are only slightly related. This fact makes the estimating of reliability very difficult. Fortunately, the performance of candidates on each subtest is so high and the standard error of measurement is so low, that very few candidates are in jeopardy regarding their pass/fail status. Validity evidence supports the use of the test for making pass/fail decisions. A recommendation is to improve documentation of validity evidence. Another recommendation is to lengthen one of the subtests to achieve higher reliability.

Endodontics Test

Based on the validity evidence reviewed, this test appears well validated for test score interpretations and pass/fail decisions. The content basis for the exam, item quality, administration procedures, examiner training and validation, standard setting and comparability, and examiner consistency and reliability all are very good. Some improvement in rater consistency can be achieved, but rater consistency is not a problem considering the high reliability achieved in the current test.

Prosthodontics Test

As with the Periodontal diagnosis test, candidate performance is very high. Thus, reliability estimates are low. However, the standard error of measurement is also very low, so only a few candidates who have scores close to cut score have some uncertainty about their pass/fail status. This test could be longer and consist of higher quality test items to improve reliability. Documentation of validity evidence should be improved.

PATP Test

This test has a wealth of validity evidence supporting it. Primary strengths are clear statements about content, sound administration, high examiner consistency, high reliability, and effective training and trainer validation. There are no threats to validity present in this test.

References

American Association of Dental Examiners (2003). *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.

Becker, D. F., & Pomplum, M. R. Technical reporting and documentation. In S. M. Downing and T. M. Haladyna (Eds). *Handbook of Test Development* (pp. 711-724). Mahwah, N.J.: Lawrence Erlbaum Associates.

Buckendahl, C., & Plake, B. S. (2006). Evaluating tests. In S. M. Downing and T. M. Haladyna (Eds). *Handbook of Test Development*, pp. 725-738. Mahwah, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1988). Five perspectives of the validity argument (pp. 3-18). In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Downing, S. M. & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,*61-82.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.) *Educational Measurement*, 4th edition, pp. 65-110. Westport, CN: Praeger.

Haladyna, T. M. (1998). *An evaluation of the Western Region Examination Board Dental Examination.* Phoenix: Author

Haladyna, T. M. (2002a). Supporting documentation: Assuring more valid test score interpretations (pp. 89-108). In J. Tindal & T. M. Haladyna (Eds.) *Large scale assessment for all students*. Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (2002b). Research to improve large scale testing. pp. 483-497. In J. Tindal & T. M. Haladyna (Eds.) *Large scale assessment programs for all students*. Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd edition). Mahwah, NJ: Lawrence Erlbaum Associates).

Haladyna, T. M. (2005). *An evaluation of the Western Region Examining Board Dental Examination*. Phoenix: Author.

Haladyna, T. M. (2006). Roles and importance of validity studies in test development. In S. M. Downing and T. M. Haladyna (Eds). *Handbook of Test Development*, pp. 739-758. Mahwah, N.J.: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Haladyna, T. M., & Hess, R. K. (1999). Conjunctive and compensatory standard setting models in high-stakes testing. *Educational Assessment, 6*(2) 129-153.

Kane, M. T. (2006a). Content-related validity evidence. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development*, pp. 131-154. Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. T. (2006b). In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Madaus, G. F. (1992). An independent auditing mechanism for testing. *Educational Measurement: Issues and Practice, 11(1)*, 26-31.

McCallin, R. (2006). Test administration. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of*

test development, pp. 625-652. Mahwah, NJ: Lawrence Erlbaum Associates.
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.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan.
Raymond, M. & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing and T. M. Haladyna (Eds). *Handbook of Test Development*, pp. 181-224. Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix: Archive of Cited Documents Providing Validity Evidence

American Association of Dental Examiners (2005). *Guidance for clinical licensure examinations in dentistry*. Chicago: Author.

Baker, J. (March 14, 2003). Dentistry interactive computerized examination system for prosthetics. Test specifications. Victoria, Australia: Zoomorphix.

Baker, J. (November 11, 2004). *Specifications for data exchange between CSW exam builders and testing agency software* (CRDTS, SRTA, WREB data bases. Victoria, Australia: Zoomorphix Systems

Baker, J. (June 24, 2005). *Periodontal report*. Victoria, Australia: Zoomorphix.

CSW (January 29, 2005). *CSW Joint Prostodontic Committee minutes*. Phoenix

CSW (September 9-10, 2005). *Periodontal Committee Meeting Minutes*. Phoenix: Author.

CSW (October 27, 2005). *Periodontal Committee Meeting Minutes*. Phoenix: Author.

CRDTS/SRTA/WREB (November 14-15, 2003). *Joint Periodontal Committee Minutes*. Phoenix, Author.

CRDTS/SRTA/WREB (February 14-15, 2004). *Joint Periodontal Committee Minutes*. Phoenix: Author.

CRDTS/SRTA/WREB (March 27-28, 2004). *Joint Periodontal Committee Minutes*. Phoenix: Author.

CRDTS/SRTA/WREB (July 10-11, 2004) *Joint Periodontal Committee Minutes*. Phoenix: Author.

CRDTS/SRTA/WREB (January 28-29, 2005) *Joint Periodontal Committee Minutes.* Phoenix: Author.

CRDTS/SRTA/WREB (June 24-25, 2005). *Joint Periodontal Committee Minutes*. Phoenix: Author.

CRDTS/SRTA/WREB (December 10, 2005). *Joint Periodontal Committee Minutes*.

Hammond, D., & Buckendahl, C. (March 2007). *Technical report for CSW computer simulations LLC.* Phoenix: CSW.

Hammond, D., & Buckendahl, C., (March 2008). *Technical report for CSW computer simulations LLC*. Phoenix: CSW

Hammond, D. (June 2009). *Technical report for CSW computer simulations* LLC. Phoenix: CSW.

Hammond, D. (February 2010). *Technical report for CSW computer simulations LLC.* Phoenix: CSW.

Ramos, D. (October 30, 2008). *Changes for the examination year 2009.* Phoenix: WREB.

Ramos, D. (June 23, 2009). *Description of scoring and standard setting*. Phoenix: WREB.

Ramos, D. (October 14, 2009a). *School coordinator preparation instructions: 2010 WREB Dental Examination*. Phoenix: WREB.

Ramos, D. (October 14, 2009b). *Examination year preparation*. Phoenix: WREB.

Ramos, D. (October 14, 2009c). *2010 school information packets*. Phoenix: WREB.

WREB (August 15-16, 1998). *Prosthetics committee meeting*. Phoenix, Author.

WREB (January 9-10, 1999). *Prosthetics committee meeting*. Phoenix, Author.

WREB (April 17, 1999). *Prosthetics committee meeting*. Phoenix, Author.

WREB (April 23, 1999). *Draft test specifications–prosthetics*. Phoenix, Author.

WREB (September 25-26, 1999). *Prosthetics committee meeting*. Phoenix: Author.

WREB October 8, 1999. *Prosthetics committee meeting*. Phoenix, Author.

WREB (December 11-12, 1999). *Prosthetics committee meeting*. Phoenix, Author.

WREB (September 25-26, 1999). *Prosthetics test specifications*. Phoenix: Author.

WREB (August 18-19, 2001). *Prosthetics committee meeting*. Phoenix: Author.
WREB (November, 10-11, 2001). *Prosthetics committee meeting*. Phoenix: Author.
WREB (February 15, 2002). *Prosthetics committee task force meeting*. Phoenix: Author
WREB (August 2-3, 2002). *Prosthetics committee meeting minutes*. Phoenix: Author.
WREB (October 19, 2002). *Prosthetics committee meeting minutes*. Phoenix: Author.
WREB (January 11, 2003). *Western Regional Examining Board By Laws* (As amended by the Membership. Phoenix: Author.
WREB (February 1-2, 2003) *Periodontal committee meeting minutes*. Phoenix: Author.
WREB (February 14, 2003). *Prosthetics committee meeting minutes*. Phoenix: Author.
WREB (July 25-27, 2003). *Prosthetics committee meeting minutes*. Phoenix: Author.
WREB (2005). *An Evaluation of the Western Regional Examining Board's 2005 Dental and Dental Hygiene Examination Program.* Phoenix, Author
WREB (July 5, 2005). *WREB board of directors meeting minutes*. Phoenix: Author.
WREB (2006). *An Evaluation of the Western Regional Examining Board's 2005 Dental and Dental Hygiene Examination Program.* Phoenix, Author
WREB (January 7, 2006). *WREB board of directors meeting minutes*. Phoenix: Author
WREB (May 19-20, 2006). *Operative test committee meeting minutes*. Phoenix: Author.
WREB (July 19, 2006). *Examination review committee meeting minutes*. Phoenix: Author.
WREB (July 20, 2006). *WREB board of directors meeting minutes*. Phoenix: Author.
WREB (September 8-9, 2006). *Operative test committee meeting minutes*. Phoenix: Author.
WREB (2007). *An Evaluation of the Western Regional Examining Board's 2005 Dental and Dental Hygiene Examination Program.* Phoenix, Author
WREB (January 7, 2007a). *Amendment and restatement of bylaws of WREB*. Phoenix: Author.
WREB (January 7, 2007b). *WREB board of directors meeting minutes*. Phoenix: Author
WREB (February 2-3, 2007). *Operative test committee meeting minutes*. Phoenix: Author.
WREB (May 15, 2007). *Practice analysis for general dentist*. Phoenix: Author.
WREB (July 11, 2007). *Examination review committee meeting minutes*. Phoenix: Author.
WREB (July 12, 2007). *WREB board of directors meeting minutes*. Phoenix: Author.
WREB (September 7-9, 2007). *Operative test committee meeting minutes*. Phoenix: Author.
WREB (November 30-December 1, 2007). *Operative test committee meeting minutes*. Phoenix: Author.
WREB (Summer 2008). *Newsletter*. Phoenix: Author.
WREB (Fall 2008). *Dental student newsletter*. Phoenix: Author.
WREB (January 5, 2008). *WREB board of directors meeting minutes*. Phoenix: Author.
WREB (February 29-March 1, 2008). *Operative test committee meeting minutes*. Phoenix: Author.
WREB (May 9-10, 2008). *Operative test committee meeting minutes*. Phoenix: Author.
WREB (July 8, 2008). *Proposed scoring detail*. Phoenix: Author.
WREB (July 9, 2008). *Examination review committee meeting minutes.* Phoenix: Author.
WREB (July 10, 2008). *WREB board of directors meeting minutes*. Phoenix: Author.
WREB (2009). *WREB Annual report*. Phoenix: Author.
WREB (March 6, 2009). *Operative test committee meeting minutes*. Phoenix: Author.
WREB (April 9, 2009). *Periodontal Treatment test committee meeting minutes*. Phoenix: Author.
WREB (May 15-16, 2009). *Endodontics committee meeting minutes*. Phoenix: Author.
WREB (July, 15, 2009). *Examination review committee meeting minutes.* Phoenix: Author.
WREB July 16, 2009). *WREB board of directors meeting minutes*. Phoenix: Author.

WREB (September 11-12, 2009). *Operative test committee meeting minutes*. Phoenix: Author.
WREB (2009a). *Criteria to become a WREB examiner*. Phoenix: Author.
WREB (2009b). *Dental candidate guide*. Phoenix: Author
WREB (2009c). *Examiner manual*. Phoenix: Author.
WREB (2009d). *Policy and procedures manual*. Phoenix: Author.
WREB (Winter/spring 2009). *Newsletter*. Phoenix, Author.
WREB (Fall 2009a). *Dental student newsletter*. Phoenix: Author.
WREB (Fall 2009b). *Newsletter*. Phoenix: Author.
WREB (2010a). *Examiner standardization*. Phoenix: Author.
WREB (2010b). *Annual report for 2009*. Phoenix: Author.
WREB (January 30, 2010). *WREB board of directors meeting minutes*. Phoenix: Author.
WREB (February 2010). *An evaluation of the Western Regional Examining Board's 2008 dental and dental hygiene examination program*. Phoenix: Author.