An Evaluation of the

Western Region Examining Board

Dental Hygiene Examination

Dr. Thomas M. Haladyna
Professor Emeritus
Arizona State University
tmh@asu.edu

March 11, 2010

Acknowledgments

The evaluation of any examination program is complex and challenging. As with the previous evaluations of WREB's *Dental Hygiene Examination*, 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. Robin Krych also provided many documents that aided my review. Radley Masinelli provided excellent information about WREB's security policies and procedures. WREB's board is commended for undertaking this evaluation. I am solely responsible for any errors or lacks of clarity or inaccuracies in this document.

Dr. Thomas M. Haladyna
Phoenix, Arizona
March 2010

Table of Contents

INTRODUCTION

Examining boards like WREB periodically undergo external evaluation to determine how validly test scores are interpreted and used by participating states. The process of evaluation entails considerable study of documents and some data analysis. This evaluation report has seven parts:

- Part I addresses why this evaluation is being conducted.
- Part II describes the *Dental Hygiene Examination*.
- Part III discusses validity and the investigative process known as *validation*.
- Part IV identifies professional testing standards that apply to this test.
- Part V briefly discusses the topic of legal defensibility.
- Part VI reports the validity evidence collected to support validity.
- Part VII is a summative evaluation.

References are provided at the end of this report. The appendix shows documents reviewed for this evaluation and are part of this validity evidence.

PART I: WHY IS THE *DENTAL HYGIENE 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 over these years. WREB's corporate office is in Phoenix, Arizona. Its bylaws were amended by its membership (WREB, January 11, 2003; January 2007). A history of WREB is available on its website (wreb.org/Information/History.htm). Another good source is any annual report (WREB, 2008).

WREB provides information about the competency of candidates for licensure as dental hygienists to 14 member states and other participating states. This information is used with other information to decide licensure for each candidate in a state. WREB also sponsors a *Dental Examination*, an *Anesthesia Examination*, and a *Restorative Examination*.

WREB has a Board of Directors (also known as the *governing board*) and an Examination Review Committee that oversees the Dental Hygiene Examination. WREB has a Dental Hygiene Subcommittee that meets regularly, reviews policies and procedures, and recommends changes intended to improve the examination. The structure of subcommittees and the way staff serves WREB and the committees are clearly shown in its annual report (WREB, 2008). WREB's Examination Review Committee also issues annual reports at the Board of Directors meetings (WREB, July 21, 2005; July 20, 2006; July 12, 2007; July 10, 2008; July 16, 2009).

Responsibilities of Examining Boards Like WREB

Examining boards provide important information to states concerning candidates for licensure 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.

The main concern of any examining board is to increase the likelihood that the professionally licensed person will treat the public that they serve safely. The content of these examinations is professional competence. This content usually consists of knowledge, skills, and abilities (KSAs). Specifying 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. The examination alone does not determine which candidate receives a license. However, in most states and jurisdictions, passing an examination is the most important criterion for licensure that all candidates must achieve if they are to be allowed to practice.

Evaluation of a 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 evaluations is to verify that a test is providing valid information about the professional competency of its candidates. Such evaluations also provide constructive criticism that may improve validity.

Every test consists undergoes three important, logical, sequential, related steps:

1. defining of a profession in terms of 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 system for validation (Kane, 2006a). One might think of validation as an investigation bearing on validity. Because no examination or battery of tests is completely adequate for measuring competence and because no system of making pass/fail decisions is infallible, validation serves two very useful purposes: (1) It determine how valid test score interpretations and uses are in the opinion of the evaluator, and (2) the evaluator offers constructive criticism aimed at improving the test and validity.

Two earlier evaluations of WREB's *Dental Hygiene Examination* provide validity evidence and opinions that were current to the date of each evaluation's publication (Haladyna, 1998; 2005). The current report updates this validity evidence. The organization and emphases in the current report differ slightly from past reports to reflect changes in the concept of validity and validation (see Kane, 2006a, 2006b). Greater emphasis is placed on reliability in this evaluation.

PART II: DESCRIPTION OF THE *DENTAL HYGIENE EXAMINATION*

The *Dental Hygiene Examination* provides examination scores to states for use in making licensing decisions for dental hygienists. Table 1 provides highlights of this test. More detailed description appears in the *Dental Hygiene Examination 2010 Candidate Guide* (WREB, 2010a).

Table 1: Highlights of the *Dental Hygiene Examination*

The examination consists of four parts:

1. Probe Depths/Recession–15 points
2. Extra/Intraoral Exam–10 points
3. Calculus Removal and Tissue Trauma–75 points
4. Penalties (points deducted for infractions)

Total examination score is 100 points.

Possible Point Deductions
x-ray penalty–4 points
First patient unacceptable–4 points
Second patient unacceptable–3 points
No acceptable patient–failure of the examination
Late penalties–1 point per minute that patient is late for checkout.

Cut (passing) score is 75.

Information about validity can be obtained from annual technical reports and other documents in the archive and from previous evaluations (Haladyna, 1998, 2005; WREB, February 2010). This report provides references to many documents 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. Data bearing on this threat to validity is presented in this report.

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

This 100-point scale is not a raw-score scale. Any multiple-choice examination lends itself to raw score and percentage correct scores, but a performance examination has different scoring rules. This examination includes objective categorical judgments or the use of rating scales in combination to form a total score. Professional judgment is a key element in determining these scoring rules. Although such judgments are often subjective, the use of subject-matter experts (SMEs) is validated if a consensus is reached about the evaluation criteria and consistency of judgments is empirically evidenced. These SMEs must be highly qualified. WREB has developed extensive criteria for appointing SMEs (WREB 2009a). 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 *2010 Dental Hygiene Candidate Guide*.

PART III: VALIDITY

As noted in the introduction, the most important concern in any examination is ***validity***. An examination score should provide a valid interpretation of a candidate's professional competence. If an examination score is used as one criterion to advance or prevent advancement of a candidate to licensure, the decision to pass or fail this candidate must also be valid. Therefore, the focus of this evaluation is validity. All other ideas about test quality are subsumed under validity.

Validity involves the professional judgment of the reasonableness of an interpretation or use of an examination score. The *Standards for Educational and Psychological Testing* (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. These guidelines were applied in this evaluation.

What does an examination score obtained from the *Dental Hygiene Examination* mean? How valid is it for a state to make a pass/fail decision based on this examination score? Thus, validity does not address an examination, so the term *examination validity* is inappropriate. Validity focuses on the meaningfulness of an interpretation 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 dental hygienist competence 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 lays out what we plan to measure and how the measure will be validly interpreted and used;

2. a claim that the measure is 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 supporting the claim. Negative evidence should be inconsequential. Negative evidence leads to recommendations to study, assess, and eliminate or reduce the factors causing this negative evidence. By that, validity is increased.

Table 2 on the next page shows the constituent elements in validation, which is the process of obtaining evidence supporting the claim about validity. This table also shows the reasoning process used in this validation.

Table 2: Validation of WREB's *Dental Hygiene Examination*	
Argument	The American Dental Association administers a *National Board Dental Hygiene Examination*. This examination measures the knowledge and skills thought to be necessary for safe and competent dental hygiene practice. This examination derives principally from a practice analysis of the profession of dental hygienists. The WREB *Dental Hygiene Examination* is a clinical performance examination intended to directly measure dental hygiene clinical competence. These two examinations represent complementary aspects of dental hygiene competence. WREB's *Dental Hygiene Examination* is the capstone in this licensing process for dental hygienists.
Claim About Validity	WREB claims that examination scores obtained from candidates represent dental hygiene 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 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	One finds in this report that there are no gaps in the validity evidence sought.
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

To organize the evaluation of validity, we have categories of validity evidence that include examination content, item quality, reliability, examination administration, training of examiners and scoring, comparability, standard setting, reporting, candidate guide, and security. This body of evidence is evaluated holistically, not individually. However, a weak link in the body of evidence is serious and should be investigated further. Validity studies look into weaker

evidence with the goal of improving this evidence. Part VI of this report presents this validity evidence. This evidence includes recommended procedures, documentation, and statistical analyses. This evidence is used in the same manner that a jury weighs evidence and decides that supports either the prosecutor's claim or the defense's claim.

Evidence Weakening the Argument

No examination reaches its ultimate in validity. All examinations undergo improvements in validity in an evolutionary path, but the road is steep and long. In any evaluation, honest examination of evidence that undermines validity is seldom done by examination sponsors (Cronbach, 1988). In this evaluation, evidence undermining validity was sought.

According to Messick (1989), two kinds of evidence that weaken validity are construct under-representation (CUR) and construct-irrelevant variance (CIV). Construct is another name for the domain of KSAs that comprise dental hygiene 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 is only as strong as its weakest link.

CUR is present if the definition of dental hygienists clinical competence does not match what the examination measures. ***Fidelity*** is the technical term we use to assess the connection of the tasks on the examination to the definition of competence for dental hygienists. If we used a multiple-choice examination of scientific knowledge or a multiple-choice examination of professional knowledge, we would not be representing dental hygiene clinical competence adequately. That is why the *National Board Dental Hygiene Examination* is a necessary licensing requirement but it is not sufficient. These multiple-choice tests under-represent the construct of dental hygiene competence. When we combine the results of the *National Board's Dental Hygiene Examination* with WREB's *Dental Hygiene Examination*, we have important complementary pieces of information that provide adequate representation of the construct of dental hygiene competence. Thus, participating states use both the National Board's and WREB's examinations due to their complementary nature.

Summary

This section on validity is best summarized in Table 2. It shows that we start with a definition of dental hygiene competence, then formulate an argument about the validity of using WREB's *Dental Hygiene Examination* scores as a measure of clinical competence. A claim is made by WREB for its client states that using these examination scores in that way is valid. In this report, evidence was collected and displayed both supporting and weakening this claim for validity. After all evidence is collected and assessed, a summative judgment is made about the validity of WREB's examination score interpretation and use. Participating states use both the National Board and WREB's examinations to decide whether a candidate receives a license to practice in that state. All licensing authorities have a responsibility to the public to do this. The National Board and 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. All of the referenced guidelines bear on the overall judgment of validity. A set of new standards is being developed, but these new 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 terms of principles related to validity.

Table 3 on the next page summarizes some more important standards 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 of validity evidence. WREB's annual technical report is one source of documentation. This report is another source. WREB keeps an archive of documents that bear on validity. 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 information about the importance of documentation includes Becker and Pomplum (2006) and Haladyna (2002).

Chapter 7: Fairness. As this examination is used in licensing dental hygienists, the issue of fairness is an important one. The design and administration of the *Dental Hygiene Examination* do not in any way violate any standard of fairness discussed in chapter 7. Examiners have no contact with candidates, and only see their 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 Hygiene Examination*.

Chapter 9: Linguistic background. As this clinical performance examination involves patient treatment under simulated natural conditions involving patient-dental hygienist interaction, no threat due to inadequate linguistic background is perceived. Most of the candidates are trained in the United States and received their degree from one of the dental hygiene schools. 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. This

examination should not simplify the language to accommodate an English language learner, because part of the professional responsibility in licensure is to ensure that the licensee has sufficient verbal ability to read, write, speak, and listen in English at an appropriate level for the profession of dental hygienists.

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 2010 *Dental Hygiene Candidate Guide* (WREB, 2010a) 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. This category of standards applies to WREB's participating states who use examination information. Overall, 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 providing all participating states with such information that supports their uses of examination scores. WREB's *Dental Hygiene Candidate Guide* (WREB, 2010a) is published every year and provides much information. WREB's technical reports are good sources of information (WREB, February 2010). These reports constitute sources of information that might be made available to states. WREB's website also provides public access to these documents.

PART V: LEGAL DEFENSIBILITY

Besides providing the highest quality examination possible, WREB does not want 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 .

Validation is an effort to provide evidence that supports the examination and its purpose. By undertaking a validation, WREB provides assurance to its participating states that the examination score information can be used validly. Such validation efforts can also be used with various constituencies and the public to avoid litigation. When potential litigants know that validation has been done and the evidence is available, they are less likely to challenge the examining board.

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) provide a useful discussion of legal threats and validity.

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 board. (See http://www.wreb.org/).

PART VI: VALIDITY EVIDENCE

This section of the evaluation is the longest and most important in this report. As noted previously, a reader should not consider the evidence individually but instead collectively. The summative judgment offered at the end of this report is based on the evaluator's ability to integrate evidence and determine how valid the claim for validity is. For the purpose of constructive criticism, for each category of evidence a conclusion is drawn about its adequacy. Later in this report, the evidence is summarized and the summative evaluation is offered.

1. Content-related Validity Evidence

The most fundamental type of validity evidence for a credentialing examination is content-related (Kane, 2006b). A dental hygiene clinical examination should focus on a domain of skills and abilities thought to represent professional clinical competence in dental hygiene. A clinical dental hygiene examination should represent this domain.

A good source of guidance for identifying such test content is through a survey of the profession, known as *practice analysis* (Raymond & Neustel (2006). The focus of content-related validity evidence as discussed in the *Standards* (AERA, et al., 1999, p. 156) can be summarized in this way:

> *Often a thorough analysis is conducted of the work performed by people in the profession or occupation to document the tasks and abilities that are essential to practice. A wide variety of empirical approaches is used, including delineation, critical incidence techniques, job analysis, training needs assessments, or practice studies and surveys of practicing professionals. Panels of respected experts in the field often work in collaboration with qualified specialists in testing to define test specifications, including knowledge and skills needed for safe, effective performance, and an appropriate way of assessing that performance* (AERA, et al., 1999, p. 156).

Chapter 14 of the *Standards* (AERA, *et al.*, 1999) is devoted exclusively to standards affecting licensure examinations, such as WREB's. As stated in that source on page 157 and in this report, content-related validity evidence is the most important. Not only is an examination agency like WREB expected to define clinical competence in dental hygiene, but is also expected to show the validity of the constituent parts of competency as determined from a survey of the profession. Standards 14.8, 14.9, 14.10, 14.11, and 14.14 all address slightly different but complementary aspects of practice analysis as a basis for test specifications. The test specifications guide examination development.

Practice Analysis and Test Specifications

As practice analysis is virtually a requirement for formulating test specifications for a clinical competence examination such as this one, WREB has acted very responsibly toward this end. A practice survey was conducted by WREB in 1996 (WREB, September 3, 1996). A survey was conducted again (WREB, July 2005). A review of the minutes of the dental hygiene subcommittee shows extensive and frequent use of the practice survey results and actions aimed at improving many aspects of the content (e.g., WREB, July 18-19, 2005; September 9-11, 2005; 2006a; February 2-4, 2006; May 25-26, 2006; September 15-16, 2007; September 26-27, 2008; December 17-19, 2009). A report was made to WREB and specifically to the dental hygiene subcommittee concerning the results of the most current practice analysis (Hammond, December 2009). This information is crucial to the development of a new process of care examination discussed elsewhere in this report.

WREB's current *Dental Hygiene Examination* test specifications contain two components. Component A addresses patient assessment in three areas: (1) patient qualification, (2) extra/intraoral examination, and (3) periodontal measurements. Component B addresses patient treatment in two areas: (1) calculus removal and (2) tissue management. WREB's clinical performance examination is intended to help assess the clinical proficiency of a candidate for licensure.

WREB reviewed the dental hygiene examination specifications (WREB, October 15-17, 1999; June 2009). Embedded in this review was a point deduction that is part of the standard setting process for this examination. These examination specifications were officially adopted. WREB reviewed these test specifications in 2001 and considered a conjunctive standard setting strategy, which is discussed elsewhere in this report (WREB, 2001). The decision was made not to implement a conjunctive strategy due to logistical issues and substantive validity issues, including lower reliability for any subtest that would be used for a pass/fail decision. WREB issued the current version of the examination specifications (June, 2009). That document shows the point allocations and the deductions for unacceptable patient, radiography penalty, and late check-in and checkout. This information about penalties is in the *2010 Dental Hygiene Examination Candidate Guide* (WREB 2010a). The dental hygiene subcommittee also reviews the test specifications several times each year at periodic meetings (WREB, July 18-19, 2005; September 9-11, 2005; 2006a; February 2-4, 2006; September 26-27, 2008; May 25-26, 2006; September 15-16, 2007; June 4-6, 2009). A review of the minutes of these meetings will show that continuous improvements are made to increase validity.

The most significant outcome of these frequent meetings is the identification, planning, and development of a new computer-simulation examination that replaces the current intra/extra oral assessment replaces extra-intra oral examination and expands the scope of content (WREB, January 28-29, 2006; August 5-6, 2006; November 17-18, 2006; April 14-15, 2007; October 27, 2007; April 18-19, 2008; August 5, 2006; January 11-12, 2008; July 8, 2008; November 7-9, 2008; August 8-9, 2009; July 13-14, 2009; November 20-22, 2009). These documents provide evidence of a very careful, directed development of a change in the *Dental Hygiene Examination*. Although validity evidence is very good for the current examination, WREB thinks this new process of care assessment will more adequately represent the domain of patient problems encountered in patient assessment in the future.

Although using practice analysis results to update test specifications is critical to validity, WREB has shown initiative in reviewing other related documents to ensure that the content of its examination is appropriate. These documents include the *Accreditation standards for dental hygiene education programs* (ADA, 2000), *Competencies for entry into the profession of dental hygiene* (ADEA 2004), *State and community models for improving access to dental care for the under served* (ADA, October 2004), and *advanced dental hygiene practitioner draft curriculum* (ADHA, June 2006b). Thus, there is a record of continuous review of content and updating of the test specifications to stay current in the profession and, also, to take advantage of new technology to enable better assessment of candidates.

To summarize, WREB has used the results of several practice analyses and other relevant documents to update its test specifications in appropriate ways continually. Documentation was

provided supporting this assertion.

Structural Evidence

Many tests of competence have a single dimension. However, tests of clinical competence are likely to have aspects of competence that are moderately related but complementary in defining competence. This complementary nature seems true for dental hygiene. As the test consists of three parts, knowing more about the underlying structure is important because it helps to define better dental hygiene and has significant implications for how reliability is estimated. As reliability is a very significant type of validity evidence, a study of structure is crucial to estimating reliability accurately.

The distribution of scores on any examination may be normal or slightly skewed. The *Dental Hygiene Examination* is a criterion-referenced, professional competence, performance test. As such, candidates are expected to perform at a very high level. Each error is potentially harmful to patients. Because of this fact, the distribution of test scores is highly negatively skewed. That is because the candidates for licensure for dental hygiene are also highly qualified and well trained. The examination is designed to detect rare errors. Because of the high performance on the three parts of the dental hygiene examination, the medians for all scoring variables except total score fall at the upper end of their respective scales. Table 4 presents descriptive statistics for test scores for 1,478 candidates for the year 2009 who had scores ranging from 27.5% to 100%.

Table 4
Descriptive Statistics for the Total and Parts of the Examination

Scoring Variable	Minimum	Maximum	Mean	Median	Standard Deviation
Total	0	100	90.9	93.8	9.7
Probe Points	0	15	14.1	15.0	1.8
Extra/Intra Oral	0	10	8.9	9.0	1.3
Calculus	15	75	69.4	75.0	8.7

As this table shows, performance in all three parts of the examination is well above the cut score (75). The fact that the median is the maximum points possible shows that the distributions are highly negatively skewed.

Correlations Among Scoring Variables

Because these tests are designed to be criterion-referenced, norm-referenced analyses do not work very well. Correlations and reliability estimates that depend on test score variation tend to be lower-than-expected. This result is not because of inadequacies in the examination but due to the nature of the sample–most candidates are very competent. This fact will surface in other analyses and mitigate some findings. In each instance, reference will be made to this fact and will be interpreted considering this restriction in the range of scores. If many low-scoring candidates were included in the sample, the results would be more suitable for traditional norm-referenced analyses.

As noted in Table 5 below, the three parts of the examination are not correlated. As most scores are in the high 90s (on the percentage scale), this restriction of range attenuates correlation. Another interpretation is that the three parts of the examination are very independent. However, without low-performing candidates, this interpretation cannot be confirmed.

Table 5: Correlations Among Subscores of the Dental hygiene Examination

Probe Depths and Extra/Intra Oral	0.022
Probe Depths and Calculus	0.065
Extra/Intra Oral and Calculus	0.027

Probe depths and recession. No analysis of structure was done due to the unique nature of these data. Each candidate's patient undergoes 108 separate evaluations (36 locations and three examiners per location). The number of errors detected is very small. Thus, an analysis of structure would be uninformative, as conventional statistical analysis is useful when there is sufficient variation. In this instance, very little variation was observed.

Extra/intraoral. A confirmatory factor analysis with equamax rotation produced three unique factors. Table 6 lists the factors with appropriate descriptive statistics. The first factor is general intra/extra oral assessment, which has a very high mean (97%). The second factor is the eighth observation (occlusal assessment, which has a mean of 0.77 (77%). The third factor is the ninth and last observation, which has a mean of 1.30 (65%). The differences may be in terms of performance. The last two observations are much lower than the former seven observations. Correlations among these three are very low (ranging from 0.004 to 0.143).

Calculus removal. These scores were either error or no error. The mean of the 36 observations (12 areas and three examiners) varied between 0.085 and 0.114 errors. As these are error rates, all these observations are positively skewed. As with the other variables conventional methods of study of structure are uninformative.

Table 6: Results of Factor Analysis

Item	Patient Evaluation	Periodontal	Occlusal
Head and neck	0.722	-0.042	0.105
Lymph nodes	0.778	-0.086	0.161
TMJ	0.72	-0.032	0.189
Floor/mouth	0.761	-0.014	0.101
Oral Mucosa	0.601	0.089	0.08
Pharynx	0.733	0.024	0.11
Tongue	0.735	0.097	0.006
Occlusal	-0.003	0.009	0.994
Periodontal	0.001	0.989	0.008

"Variance" Explained by Rotated Components

Patient Eval.	Periodontal	Occlusal
3.661	1.006	1.089

Percent of Total Variance Explained

Patient Eval.	Periodontal	Occlusal
40.680	11.180	12.098

	Patient Evaluation	Occlusal	Periodontal
Number	1478	1478	1478
Mean	6.79	0.77	1.30
Median	7.00	1.00	2.0
Maximum	7.00	1	2.0
Minimum	0.00	0.00	0.00
Standard Dev.	0.67	0.42	0.95

Conclusion

Evidence has been presented here for appropriate development of test specifications arising from practice analysis. WREB has frequently updated content using various sources of information about content of professional dental hygiene practice. The structure of the data was not well determined due to the criterion-referenced nature of test scores. The highly, negatively skewed distributions led to low correlations, so knowing the structure of the three parts of the examination is difficult. However, the intra/extra oral assessment appeared to have three parts. This fact may affect the estimation of reliability. Overall, the content-related validity evidence reviewed and summarized in this section appears to be very strong.

2. Item Quality

Another primary type of validity evidence is item quality. Items used on the *Dental Hygiene Examination* should be developed in conformance with the definition of competence and the practice analysis that identifies the KSAs required. Items should undergo systematic development that depends on the expertise of WREB's SMEs. This process has been described as ***item validation*** (Haladyna, 2004), because the item undergoes the same procedure of validation applied to the test scores. Thus, the evidence needed to conclude that the items used in this examination have been validated include the following.

1. Practice analysis identified the KSAs needed to practice safely and competently.
2. Test specifications are created that explicate this content.
3. Items are developed to match the test specifications.
4. Items undergo intensive review by SMEs on content subcommittees.
5. The scoring protocol is assigned a point value and a procedure for arriving at each point value by the SMEs.
6. The item and the scoring protocol is field tested to assure its ability to discriminate between high- and low-performing candidates.
8. Most important, these items should have high fidelity with the criterion behavior intended–actual dental hygiene practice. There is no simulation or approximation in the *Dental Hygiene Examination*. Each subtest *directly* measures skills that qualified, competent dental hygienists perform in their professional practice.

Fidelity

Tasks on clinical examinations such as WREB should resemble those tasks performed by dental hygienists in practice. If the tasks possess fidelity with criterion behavior, part of the validity argument is that the content of the *Dental Hygiene Examination* has high fidelity with the tasks performed by dental hygienists in practice. A review of these tasks and prior committee activities supports the fidelity argument.

Probe Depths and Recession (15 Points)

These items comprise a series of observations by the candidate scored by three examiners. Validation provides that at least two of the three examiners agree that candidates are correct or incorrect. In a conventional analysis of test item performance, difficulty and discrimination are often computed for each scorable unit (test item/task). Validity evidence for this scorable part of the test at the item (observation) level is based on the judgment of examiners that the quadrant selected for these observations is representative of the entire dentition. As these candidates' observations are exactly what a dental hygienist does with a patient, the fidelity of these items is unquestionable. The degree to which examiners agree with the candidate is reported in the section on reliability.

These items comprise a series of 36 observations by the candidate validated by three examiners. Validation provides that at least two of the three examiners agree, and where a rating scale is used, the median, not the mean is used. The mean is a biased measure of central

tendency in skewed distributions; whereas the median is the more appropriate measure of central tendency. In a conventional analysis of test item performance, difficulty and discrimination are often computed for each scorable unit (test item). Validity evidence for this scorable part of the test at the item (observation) level is based on the judgment of examiners that the quadrant selected for these observations is representative of the entire dentition. As these candidates' observations are exactly what a dental hygienist does with a patient, the fidelity of these items is unquestionable. The degree to which examiners agree with the candidate is reported in the section on reliability.

Extra/intraoral (10 points)

These nine categories of observation were identified in the practice analysis as essential aspects of dental hygiene competence. The ninth category, periodontal, receives twice the weight of other items (two points) as determined by SMEs. The items were chosen based on the recommendation of the Dental Hygiene Examination Subcommittee (WREB, December 18-19, 1998). Recent practice analyses have resulted in no changes in the examination protocol for extra/intraoral examination (Hammond, December 2009; WREB, November 2006).

Correlations of the total score for each of these categories with the total extra/intraoral score were consistently high. These coefficients ranged from 0.464 to 0.756 with a median value of 0.730. However, the last three items have the lowest of these indexes, because the error rates were greatest for these three items. A very high degree of examiner agreement was observed for the first six items. In the traditional sense of item analysis, these coefficients represent discrimination indexes and have a direct relation to the reliability estimate reported elsewhere. Despite the high degree of skewness in these scores, these coefficients are very high, which is also reflected in the reliability estimate reported in the next section.

Calculus Removal (75 points)

As with the other two subtests, the items included in this subtest are criterion behaviors of dental hygienists. Three examiners must examine the patient after treatment to decide if the calculus removal has been successful and tissue trauma has been avoided.

The 12 surfaces observed for calculus removal and tissue trauma all had low to moderate correlations with the criterion total score. These coefficients ranged from 0.385 to 0.522. The judged quality of these surfaces as observation opportunities for scoring calculus removal and tissue trauma is offered by the Dental Hygiene Examination Subcommittee (WREB, July 2000; July 18-19, 2001; July 5-7, 2002). More recent practice analysis reaffirms the decision to continue to use the calculus removal subtest (Hammond, December 2009; WREB, November 2006).

Conclusion

The items were chosen based on high fidelity with the criterion behaviors identified as necessary for dental hygienists. Item analysis does not serve any useful purposes here, because the items cannot be modified or eliminated due to each item's relationship to the definition of

competence and its fidelity to criterion behavior. The rationales for the observations used to score is a very important source of evidence for item quality. Strong evidence was noted for consistency in evaluating performance and the documentation of item development and of the rationale for item selection is extensive.

3. Reliability

Reliability is a theoretical concept that centers on two important components: the candidate's true score and random error–which can be large or small and positive or negative. Random error is unknown. Because of this, reliability can never be known; it can only be estimated. In test development, one important objective is to create a condition for high reliability to reduce random error. Candidates whose scores are close to the cut score are in danger of being misclassified due to random error. Thus, we should try to reduce random error.

We have a way of standardizing error; it is a statistic called the *standard error of measurement*. This statistic is used to help us understand the risk involved with pass/fail decisions at or near the cut score (75). Naturally, when reliability is high, the standard error is low. However, an argument about these candidates' danger of being misclassified is that any candidate scoring near the cut score is not performing at a high level. We would not want to pass a candidate with a true score below 75 and we would not want to fail a candidate with a true score above 75 due to random error. Nevertheless, if the test measures clinical competency, a fair inference is that candidates whose scores are close to the cut score have minimal competence. The way a cut score is used to make pass/fail decisions is usually based on board policy. The governing board should decide which kind of error is most serious and take steps to prevent one at the expense of the other. Or the governing board can simply set the cut score and not concern itself with borderline candidates. WREB has taken important steps to increase observations and rater consistency to ensure that reliability is very high, and the degree of random error is small, so that only some candidates are in jeopardy of being misclassified. Borderline candidates might be referred for remedial education and then retested to afford them equal opportunity to succeed.

In this section, several topics are presented that bear on reliability. First, examiner consistency is discussed. Examiners must observe candidate performance agreeably. If differences exist, then confidence is lost in the scoring. More important, reliability is lessened. The higher the number of examiners ratings per candidate, the higher the reliability. Second, the reliability of each subtest is estimated. Third, the reliability of total test score is estimated. Finally, standard error of measurement and the conditional standard error of measurement are discussed. These statistics are important in showing how many candidates' scores approach the cut score where a certain degree of uncertainty exists about their true pass/fail status. Because all tests are fallible, test scores contain a certain unknown degree of random error that may affect decision making. Increasing reliability reduces this band of uncertainty.

Examiner Consistency

Probe depths/recession. The rater consistency index consisted of the number of correct observations. For each set of three ratings, seven possibilities existed (000, 001, 010, 011, 100, 101, 110, 111). For ratings 000 and 111, we have three agreements. For all others, there is one disagreement and two agreements. With three judges the lowest possible degree of agreement is 66.7% and the highest degree of agreement is 100%. For the observations with three judges per observation, agreement ranged from a low of 97.7% to 99.7%. The mean percent of agreement was 98.4%. For this part of the test examiner consistency is extremely high.

Extra/intra oral examination. The extra/intra oral subtest has a different structure than the probe depths/recession subtest. A six-point rating scale was used. Therefore, it is possible to compute agreement indexes using coefficient alpha, which is appropriate for ratings that are unidimensional. Table 7 below presents these examiner consistency indexes for seven of the nine scales of the extra/intra oral subtest. As the table shows, the indexes range from 0.789 to 0.823. Note the preponderance of ratings on this subtest of the *Dental Hygiene Examination.* All the means are near the ceiling of the scale, and for every one of the areas reported below, the median value is the maximum of the rating scale, 5.000. If the candidate pool had included low-scoring candidates, these rater agreement indexes would be much higher.

Table 7
Indexes of Rater Agreement for Intra/Extra Oral Assessment

Head/neck	Lymphatic	TMJ	Floor/mouth	Oral mucosa	Pharynx	Tongue
0.806	0.823	0.867	0.855	0.789	0.81	0.823

The last two parts of this subtest use a scoring system similar to the scoring method used with the probe depths/recession, so the degree of examiner consistency is reported. For the occlusal rating, the degree of rater consistency was 89.9%. For the periodontal rating (worth two points instead of one point), the degree of examiner consistency was 89.7%.

As with the previous subtest of the *Dental Hygiene Examination*, examiner consistency is extremely high. The structure study previously reported showed that these last two items are actually independent subtests. So the intra/extra oral assessment consists of three factors and is not unidimensional. This fact has implications for estimating reliability.

Calculus removal. This part of the examination has the greatest consequence on the candidate's score, and, therefore, is the most critical. To review, the candidate's score is based on 12 observations. Three examiners judge each observation. As with the other two subtests, the performance of candidates is expected to be very high, and it is. For 75 points, the agreement among raters was 93.0% with the effective range of this agreement index being 66.7% to 100%. Like other parts of the examination, rater consistency is very high.

Penalty points. The fourth part of the scoring method for the *Dental Hygiene Examination* involve deductions for inappropriate or undesirable candidate actions or consequences. WREB has improved examiner consistency of penalty deductions. Thus, reporting this finding here seems reasonable. Six categories of penalty points were assessed by examiners. The results varied between 94.7% agreement and 100% agreement in two instances with the median being 99.4%. Thus, examiner consistency is as high as or higher than other scoring categories. Some penalties were objectively assessed and were not included in this report because no judgment involved. In fact, the degree of rater consistency seems so high that the scoring category of penalty points seems objective, and the reliability of this part of the examination seems maximal. Penalty points were not considered in estimating reliability.

Reliability

Two kinds of reliability are considered here: subtest reliability and total test score reliability. The two types are related as the first contributes to the second in a complex way that will be explained in this section.

Subtest Reliability. Estimating subtest reliability is a first step in estimating total score reliability. If a conjunctive scoring model were being used, subtest reliability would be a critical issue. Because the scoring model is compensatory, subtest reliability should be reasonably high but does not have to attain the same level if a conjunctive scoring model were being used. These results below are again mitigated by the fact the distributions of scores for all three subtests are highly skewed due to a restriction in range of scores. This restriction attenuates these reliability estimates.

When a test or subtest is considered to be unidimensional (one factor), coefficient alpha is the appropriate way to estimate reliability. Where it can be shown that a test or subtest has more than one factor (multidimensional), stratified alpha is the more appropriate way to estimate alpha (Haertel, 2006, p. 77). Internal consistency (alpha) reliability estimates were computed for the three subtests with these results. For the probe depths/recession subtest, using coefficient alpha all 108 ratings (36 observations, three examiners per observation), the coefficient was 0.823. For the intra/extra oral assessment, using coefficient alpha and all 27 ratings (nine items and three assessments per item by the examiners), the coefficient was 0.874. For the calculus removal, with 12 individual observations viewed by three examiners, there is a total of 36 observations of this skill. The coefficient was 0.821. Given the skewness of scores, these coefficients are very high.

Total Score Reliability

According to Haertel (2006) and Nunnally and Bernstein (1994), when a test has a diffuse content structure, coefficient alpha (a measure of internal consistency) will very likely underestimate reliability. As the *Dental Hygiene Examination* is designed to detect errors and the commission of errors is rare, the distribution of scores is very skewed. Both factors will cause an internal consistency reliability estimate to be low. The *Dental Hygiene Examination* provides ample evidence of three parts based on both content and statistical criteria. Statistically, the correlations among these parts including a factor analysis affirm this hypothesis. According to Haertel (2006, p. 77), stratified alpha is recommended and was used to estimate reliability. The result is 0.876. This coefficient is very high. Performance tests are known to have lower reliability than multiple-choice tests due to the addition of subjectivity in examiner/rater error. Attaining such a high reliability estimate is rare for a performance test, especially in the presence of highly skewed distributions of scores for the total test and three subtests that measure unique skill sets. This result is most likely due to the 171 observations per candidate on all scorable performances.

Standard Error of Measurement

As described earlier, a standard error of measurement is a useful index to gain understanding of risk involved when using a cut score, such as 75. If most of candidates scores were clustered around the cut score, the risk of misclassification would be great. If the standard error of measurement were large (reliability is low), the risk would be considerable.

Based on the reliability estimate of 0.876, the standard error for the *Dental Hygiene Examination* is 5.82. This information presented above would be lower if item response theory were used and the conditional standard error was estimated. The above information provides a simplistic estimation of about how many candidates might be misclassified. Because total score reliability is so high, the amount of risk is very low and reasonable given. Other examination s such as the *National Board Dental Hygiene Examination* have similar standard errors. (http://www.ada.org/prof/ed/testing/nbdhe/nbdhe_technical_complete.pdf)

In a technical report, WREB (2010) reports a method for estimating the *conditional standard error* around a cut score that provides more useful information about the risk/uncertainty of making decisions for candidates whose scores fall close to the cut scores. Theoretically, the size of the standard error is believed to vary along the test score scale and is not constant, as shown in this report. However, estimating the conditional standard error of measurement is very problematic. The method reported in by WREB seems to provide useful information about the margin of error around the cut score.

How this information is used is also problematic. We have no way of knowing whether random error is positive or negative, large or small. Those candidates whose scores fall within one standard error of the cut score could be misclassified. We can allow those with scores below the cut score but within one standard error of this cut score to pass, but some of these candidates will have true scores below 75. Or we can allow those with scores within one standard error above the mean to fail, but some of these will have true scores above 75. No matter what concessions a governing board makes, the misclassification of candidates around a cut score is inevitable. The comfort in this situation is that reliability of the total scores is very high and the standard error is very low.

Conclusion

Evidence has been presented for high examiner consistency for the three parts of the examination. Evidence has been presented for high subtest reliability and total score reliability. The report of a standard error or conditional standard error presents technical problems that are not easily solved by current theory or technology. WREB has achieved a high degree of reliability in this examination, and the result is a low degree of random error. Facing the fact that the distribution of scores is so heavily skewed, this achievement is remarkable. This result might be attributed to the high number of observations for each subtest.

4. Examination Administration

In order for the *Dental Hygiene Examination* to be standardized, certain conditions must be met. All candidates must have an equivalent examination experience. That means the content of the examination has to be constant, the examination score scale should be the same, the cut score must be the same, and other conditions affecting test scores must be constant. The examination must be administered in a standardized way each time.

The *2010 Dental Hygiene Examination 2010 Candidate Guide* (WREB, 2010a) gives a very good account of the many standardized features of the administration of this examination. Coordinators at any examination site are provided useful information about examination administration (WREB, 2009g). Another important document that provides extensive discussion and information about administration is the *Dental Hygiene 2009 Examiner Manual* (WREB, 2009d). This manual provides considerable information to examiners about how the examination is administered and scored. This manual has evolved over many years. Inspection of this manual reveals many quality control checks in all aspects of the examination. Other documents provide self-assessment exercises for examiners and other information intended to improve examiner performance (WREB, 2009e, 2009f, 2010b). WREB has a differentiated staff with complementary abilities that work together to achieve a smoothly run examination.

As documented in its *2010 Dental Hygiene Examination Candidate Guide (WREB 2009a)*, the *Dental Hygiene 2009 Examiner Manual* (WREB, 2009d), and in other documents in WREB's archive, WREB addresses many issues of administration that affect validity. These issues include training of administrators of the examination, advance information that is available in the *Dental Hygiene Examination 2010 Candidate Guide*, clarity of directions in this guide, conditions of testing, patient consent forms, avoiding disruptions in the examination process, test security, monitoring candidates during the examination, responding to questions of candidates, administration instructions, and time limits. The cycle of activities for each administration is well documented in this examiner manual. Documentation exists for annual attention to issues affecting the administration of this examination (e.g., WREB, February, 19-20, 2005; September 9-11, 2005; February 2-4, 2006; May 25-26, 2006; July 8, 2008; September 26-27, 2008; June 4-6, 2009; July 13-14, 2009; December 17-19, 2009).

Conclusion

The validity evidence addressing administration is described briefly above and well documented in the above references. These documents are substantial in scope. The *Standards* (AERA et al., 1999) contain 46 specific statements regarding administration. WREB is commended for dedicating so much effort and resource in standardizing examination administration. An excellent reference on test administration is provided by McCallin (2006). She identified many threats to validity arising during administration. WREB is fortunate to have a well-run examination.

5. Training of Examiners and Scoring of Candidates' Performance

The *2010 Dental Hygiene Candidate Guide* describes how examiners score performances and how their scores are converted into points (WREB, 2010a, pp. 13-14). The procedure for scoring is not a psychometric issue but depends on the expertise of the SMEs that comprises WREB's dental hygiene subcommittee that consists of SMEs. The scoring procedure for each subtest of the examination varies slightly. As noted previously, the scoring procedures make the estimating of reliability challenging because the focus is on rare error detection. Nonetheless, it has been shown that examiner consistency is very high. This section of the evaluation addresses several important aspects of training and scoring.

Selection of Examiners

WREB has an archive that documents the selection and qualifications of dental hygienists and educator dental hygienists who participate in test development and standard setting (WREB 2009a). A review of these criteria show that examiners must meet many qualifications before they are chosen as examiners. An Examiner Performance Review Committee evaluated the harshness leniency of past examiners' performances and takes action to correct any problems in examiner scoring.

Training and Evaluation of Examiners

As noted previously, WREB has extensive examiner training (see WREB 2009d). The *Dental Hygiene 2009 Examiner Manual* provides general information, procedures to follow in the candidate check-in the patient qualification, and the check out procedure. Later, after the examination is given, WREB analyzes rater performance and evaluates the examiner consistency. Many documents provide a comprehensive account of training protocols, self-assessment exercises, and procedures (WREB, 2009d, 2009e, 2009f, 2010b).

Scoring

As noted in the *2010 Dental Hygiene Examination Candidate Guide* (WREB 2010a), WREB has rules governing both scoring and deductions for various types of undesirable actions. Candidates may lose points if the patient they selected is unsuitable for examination or if other test administration conditions are not met. These scoring rules are also shown in WREB's test specifications (WREB, June 2004). These rules are subject to regular scrutiny and evaluation. For example, WREB (August 20-21, 1999) began the use of median scores instead of mean scores as an improved basis for candidate scores across three examiners' scoring.

Selection of patients. The selection of patients is a very important aspect of the examination. A candidate could fail the test by selecting patients and alternative patients who fail to qualify. The 2010 *Dental Hygiene Examination Candidate Guide* (WREB, 2010a) provides extensive discussion on pages 4-5 on patient selection. Candidates may have deductions in their score if their first and second patients are not qualified for the examination.

Probe depths/recession (15 points). Scoring is based on a point deduction method, where 2.5 points are deducted for each error up to a maximum of 15 for periodontal measurements and probing. For gingival recession, 2.5 points are deducted. The sum of deductions for this category cannot exceed 15 points. Of the total sample for 2009 for this analysis, 71.6% of the scores were 15, and 23.4% had a score of 12.5. The competency level of all candidates is very high. There is a substantial correlation between points awarded on this subtest and total score (0.924).

Extra/Intraoral (10 points). For each of the first eight categories, one point is the maximum score; for the ninth category (periodontal assessment) two is the maximum score. A six-point rating scale (0-5) is used to rate performance on the first seven categories. Category 8 (Occlusion) is scored 0-1. Category 9 is scored 0-2. A conversion table changes ratings to fractions of a point for the first seven categories (WREB, 2010a).

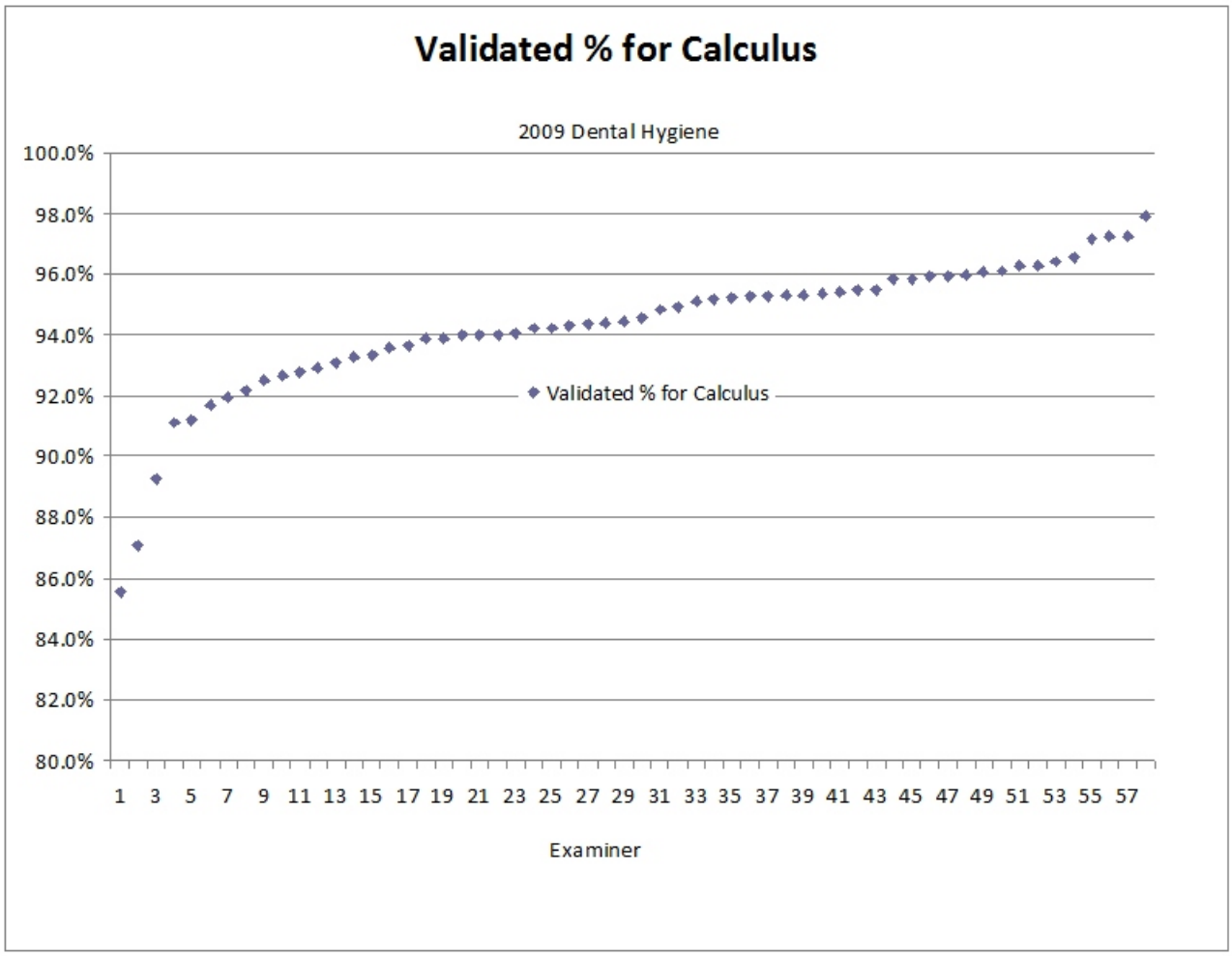
Calculus removal (75 points). This is the most consequential part of the examination. Any candidate loses six points for every validated error. A validated error is one where two or three examiners agree that the candidate has erred in removing calculus or there is an instance of tissue trauma. For the 1,512 candidates in this sample, the range of scores was 0 to 75. The mean score was 67.4 (92.5%), which is well above the cut score standard of 75 (equivalent to 56.25 for a 75-point subtest). Thus, the typical scores for calculus removal are very high.

Examiner Accuracy

Examiner consistency was discussed in the section on reliability, because consistency is directly and statistically related to reliability. The higher the examiner consistency, the higher the reliability. In this section, a family of errors is evaluated. Earlier in this report, the term CIV was defined as any factor that influences a test score but is not relevant to what is being measured. CIV needs to be reduced or eliminated, and by doing so validity is increased.

A general term for this family of errors is *rater effect*. The most serious of rater effect is severity/leniency. Some examiners are harsher than other examiners, and some examiners are more lenient than others. WREB uses a scoring method that defends against this kind of unfair rating.

Analyses were done to seek information about severity/leniency in these ratings. WREB also has a monitoring system for detecting severity. The annual technical report makes reference to this monitoring (WREB, 2010), and the *Dental Student Newsletter* (WREB, Fall 2008) also provides a more detailed account of WREB's efforts to ensure that examiners maintain a high degree of accuracy in their assessments. An examination of a typical examiner profile shows a detailed analysis of rating patterns with a severity index. Examiners are also evaluated for accuracy (Hammond, February 16, 2010). The figure on the next page shows a profile of examiner accuracy. Thus, examiners are made aware of any serious variation, and negative impacts that might affect candidates are avoided.



Quality Control

WREB has many quality control procedures, scoring checks, and security measures in test administration procedures and employee job descriptions. Some of these procedures were documented in the section on examination administration. One example is the *Examiner Checklist* (WREB 2009g), which is a five-page document that provides a very detailed listing of activities intended to maintain the standardization of the examination administration and maintain accuracy. The *Policy and Procedures Manual* (WREB, 2009h) also contains information about quality control.

In the *Examiner Manual (WREB, 2009d)*, specific guidelines are shown and used to ensure that examiners avoid factors or conditions that may question the integrity of the examination process, such as grading candidates who may be related to the examiner in some way. WREB handles this threat to validity very well by having a photo identification procedure and other safeguards that are stated in the *2010 Dental Hygiene Examination Candidate Guide (WREB, 2010a)*.

Conclusion

Evidence was presented about the accuracy of scoring. From the data presented on reliability, examiner consistency is very high. The use of validated judgments and medians instead of means is a strategy that improves the accuracy of scores. Effective training of examiners is also very important. WREB has very thorough documentation of its training of examiners and scoring procedures. *Standards* have one standard, 5.6, that addresses the integrity of scores and fraudulently obtaining a score. WREB has procedures safeguarding against that type of fraud. Eight other standards bear on examiners. These standards address such issues as selecting examiners, qualifications of examiners, training, recalibration of examiners, feedback to examiners, and dismissal of examiners. Four scoring criteria standards exist. WREB meets these standards and provides good documentation for meeting these standards in the *2010 Dental Hygiene Examination Candidate Guide* (WREB 2010a) and the *Dental Hygiene Examiner Manual* (WREB, 2009d).

Another aspect of quality control and fairness applies to any candidate whose scores are close to the cut score. Those falling near the cut score should have their score records rechecked to ensure accuracy. Candidates should be told of this action. Consequently, challenges of examination scores will more likely avoided.

6. Scaling & Comparability

A standardized performance test should have a score scale that is same for all candidates, no matter where or when each candidate is tested. The cut score also should be the same. The 100-point scale should retain the same meaning each time the examination is given. Ideally, the test provides the same level of challenge to every candidate and examiner scoring is both consistent and unbiased for every test administration. The examination items are the same each time the examination is administered. The scoring method is the same. The selection of examiners is standardized. The criteria for selecting examiners are standardized. Although examiners at each administration may vary, all receive the same training and are also calibrated before each examination. The ratings reported in this evaluation show a high degree of examiner accuracy and consistency.

Major threats to validity for this examination are examiner consistency and CIV (systematic error) in ratings, which was discussed previously. Rater consistency is very high, and WREB's training detects and combats CIV. Although such threats to validity are omnipresent, no evidence exists thus far that suggests that these threats to validity exist in the *Dental Hygiene Examination*.

Conclusion

Given the highly standardized procedures followed in the design, administration, and scoring of the examination, the evidence supports a conclusion that the examination provides an equivalent experience each time it is administered. Differences in performance from site to site vary with the competence of candidates taking the examination. The use of the same cut score (75) seems defensible with respect to psychometric theory and our testing standards.

7. Standard Setting

The passing score for the *Dental Hygiene Examination* is set at 75 on the 100 point scale. This is the passing score WREB recommends to participating states (WREB, July 17-18, 2002; July 9, 2003; July 10, 2003; Fall 2003). Whether a state has a passing standard of 70 or 75 that is set by legal statute does not matter. States normally set arbitrary cut scores as part of their statutes for credentialing examinations. Testing agencies still have the responsibility of setting a cut score that meets standards and fairly determines who is recommended for a passing or failing decision. In one subcommittee report (WREB, September 26-28, 2003), the validity of rescaling to achieve agreement with each state's statutes regarding the passing score was discussed and resolved. For most states, adjustments are made by the test sponsor so that the recommended cut score set by the governing board is at the same point as chosen arbitrarily by the state. States are advised to choose cut scores according to established standards and not arbitrarily. Test sponsors such as WREB make these adjustment in cut scores as an accommodation. In no way are candidates' pass/fail status jeopardized unfairly by such adjustments.

Passing Score Studies

WREB has periodically conducted passing score studies. For the extra/intra oral calibration exercise, (WREB, October 29, 2004), Dixon reported a procedure used by her subcommittee to recommend a passing score. WREB provides extensive documentation for issues related to setting the cut score (WREB, July 7-8, 2003). The passing score has not been changed since.

Conclusion

Documentation of procedures used to set the standard was provided. WREB meets the guidelines regarding the setting of a cut score and its documentation.

8. Reporting

The *2010 Dental Hygiene Examination Candidate Guide* (WREB, 2010a) provides description about scoring. Score reports are designed to reveal candidate performance in all aspects of the examination in a point basis against possible points to be earned. Confidentiality of candidates' results is ensured. Candidates graduating from dental hygiene 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 not 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. Score reports are confidential and are not public documents. An inspection of a typical school score report and individual score report shows clear and comprehensive candidate performance that includes previous attempts, point deductions, and performance points for each of the three parts of the examination. A pass/fail decision also appears in the score report. For the individual candidate score report, detailed information is published for the extra/intra oral assessment and for point deductions. The probe and recession and calculus and tissue trauma scores have no subscore information due to their structure, which is unidimensional.

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.

9. Candidate Guide and Rights of Test Takers

The *2010 Dental Hygiene Examination Candidate Guide* (WREB, 2010a) has been cited often in this report. This booklet contains essential information for candidates. The table of contents for this 29-page booklet provides general information, performance evaluation information, patient criteria, criteria for teeth and for calculus detection and removal, and examination procedures. The booklet is published each year and is updated as needed. Besides this guide, WREB's web page is helpful, and if candidates prefer can contact the WREB office by phone or by email. WREB's quarterly newsletters also have information that may be useful to candidates.

WREB also issues regular newsletters of general interest, of interest to dental students, and of interest to dental hygiene students (WREB, Summer, 2008; Fall, 2008; Winter/Spring, 2009; Fall, 2009a; Fall 2009b; Fall 2009c). 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.

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 *2010 Dental Hygiene Candidate Guide* (WREB, 2010a) 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 test.

10. Security

The *WREB Policy and Procedures Manual* (WREB 2009h) discusses security. Another document is *Dental Hygiene Examination Security* (WREB, 2009b). WREB has taken many steps to safeguard against cheating and other threats to validity during the examination. This aspect of its security policy and procedures is well in place.

WREB has security policies and procedures for technology hardware and software. Organization data is stored and processed on servers run from locked rooms. The server rooms are secured using keypad entry locks. These rooms are limited to executive and information technology team access only. 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 also once per week offsite. Access in and out of the WREB internal network is guarded by hardware and software fire walls. 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 IT team.

Offsite critical data is also copied for redundancy. 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.

At examination sites, electronic scoring system computers are configured with data encryption. Files are only accessible on site by limited approved personnel. In the unlikely event of a stolen device, data files are essentially useless as they remain encrypted until WREB unlocks data.

It is extremely unlikely that a single examiner or team of examiners could undermine the validity of any examination. Examiners are subject to high standards of performance and scrutiny. Self-interest or other factors may contribute to unwarranted ratings. This kind of behavior 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 minimize this possibility. The integrity of examiners is discussed in the *Dental Hygiene 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, 2009). WREB has provided excellent validity evidence bearing on security. It meets the standard 5.9 (Standards, AERA, *et al*., 1999).

VII: SUMMATIVE EVALUATION

The definition of dental hygiene practice, the argument supporting the valid interpretation of test scores, the claim for validity, and the evidence presented in this document all points to a very high quality test. Several factors contribute to this evaluation.

1. WREB has a long and consistent history of planning, developing clinical performance tests of high quality and validating test score interpretations and uses. This tradition continues. WREB's annual reports, technical reports, documentation, evaluation reports like this one contribute to a growing reputation in the United States for high quality examinations. The greatest strength is the overall commitment to excellence that permeates all aspects of the testing program. This includes the Board of Directors, the Examination Review Committee, the Dental Hygiene Examination Committee, and the staff who plan and administer the test and the participation of states, dental hygiene schools, and other constituencies that support such testing programs, such as the American Association of Dental Examiners, and the guidelines they recently published with WREB's help and support. The *Dental Hygiene Examination* Subcommittee operates effectively. They are advocates for improvement, and their minutes and recommendations give evidence of this commitment.

2. Ample evidence is presented for a systematic growth in the validity evidence needed for validation. Threats to validity are few and minor in scope. Previous reports also discuss the abundance of evidence favoring valid interpretations and uses of test scores (Haladyna, 1998; 2005).

3. The *Standards* (AERA, et al., 1999) provides many guidelines for clinical performance testing programs like WREB's to follow. The opinion of this evaluator is that these guidelines are followed, and documentation has been provided to support this opinion.

Thus, considerable validity evidence exists that supports WREB's participating states using these test scores for making pass/fail decisions for licensure in dental hygiene.

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. (1996). Model for evaluating high-stakes testing programs: Why the fox should not guard the chicken coop. *Educational Measurement: Issues and Practice, 15*, 5-12.

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 Hygiene Examination.* Phoenix: Author

Haladyna, T. M. (2002). 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. (2002). 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 Hygiene 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 .

Hammond, D. (Fall, 2003). Obsession with saving the "ideal" lesion for the WREB examination. *WREB Dental Student Newsletter*.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

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.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

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.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd edition). New York: McGraw-Hill.

Appendix: Archive of Cited Documents Providing Validity Evidence

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

American Dental Association–ADA (2000). *Accreditation standards for dental hygiene education programs*. Chicago: Author.

ADA (October 2004). *State and community models for improving access to dental care for the underserved*. Chicago: Author.

American Dental Educators Association (ADEA (July 2004). Competencies for entry into the profession of dental hygiene, *Journal of Dental Education, 68*(7), 745-749.

American Dental Hygienists Association (ADHA) (August 2005). *Practice Act overview chart.* Chicago: ADHA. Also available on its website:http://www.adha.org/.

ADHA (June 2006a). *Clinical practice guideline*. Chicago: Author

ADHA (June 2006b). *Advanced dental hygiene practitioner draft curriculum*. Chicago: Author

ADHA (March 10, 2008). *Standards for clinical dental hygiene practice.* Chicago: Author.

ADHA (2009). *ADHA Practice act overview chart*. Chicago: Author.

Hammond, D. (December 2009). *WREB dental hygiene practice analysis*. Phoenix: WREB.

Hammond, D. (February 16, 2010). Memo to examiners. Phoenix: WREB.

Joint Commission on National Dental Examinations (2009). *National Board Dental Hygiene test specifications*. Chicago: Author.

WREB (January 11, 2003). *Western Regional Examining Board By Laws* (As amended by the Membership. Phoenix: Author.

WREB (February 19-20, 2005). *Dental hygiene examination development committee*. Phoenix: Author.

WREB (July, 18-19) 2005). *Dental hygiene subcommittee: recommendations and justifications*. Phoenix: Author.

WREB July 20, 2006). *WREB Board of Directors Meeting minutes*. Phoenix: Author.

WREB (September 9-11, 2005). *Dental hygiene subcommittee meeting*. Phoenix: Author.

WREB (2006). *Report of the dental hygiene development committee*. Phoenix: Author.

WREB (January 28-29, 2006). *Dental hygiene computer simulation committee*. Phoenix: Author.

WREB (February 2-4, 2006). *Dental hygiene subcommittee meeting*. Phoenix: Author.

WREB (May 25-26, 2006). *Dental hygiene subcommittee meeting*. Phoenix: Author.

WREB. (August, 5, 2006a) *Computer-simulation test specifications comprehensive dental hygiene care.* Phoenix: Author.

WREB (August 5, 2006b). *Test specifications*. Phoenix: Author.

WREB (August 5-6, 2006. *Computer simulation committee meeting.* Phoenix: Author.

WREB (November 2006). *WREB Dental hygiene practice analysis report*. Phoenix: Author

WREB (November 17-18, 2006). *Dental hygiene ad hoc computer simulation committee*. Phoenix: Author.

WREB (January 2007). *Amendment and restatement of bylaws of WREB*. Phoenix: Author.

WREB (April 14-15, 2007). *Dental hygiene computer simulation committee.* Phoenix: Author.

WREB July 12, 2007). *WREB Board of Directors Meeting minutes*. Phoenix: Author.

WREB (October 27, 2007). *Dental hygiene computer simulation committee meeting*. Phoenix: Author.

WREB (2008). *Annual report for 2008*. Phoenix: Author.

WREB (Summer 2008). *Newsletter*. Phoenix: Author.
WREB (Fall 2008). *Dental student newsletter*. Phoenix: Author.
WREB (January 11-12, 2008). *Dental hygiene computer simulation meeting*. Phoenix: Author.
WREB (April 18-19, 2008). *Dental hygiene computer simulation committee*. Phoenix: Author.
WREB (July 8, 2008). *Dental hygiene examination review committee meeting.* Phoenix: Author.
WREB (July 10, 2008). *WREB Board of Directors Meeting minutes*. Phoenix: Author.
WREB (September 26-27, 2008). *Dental hygiene subcommittee meeting*. Phoenix: Author.
WREB (2009a). *Criteria to become a WREB examiner*. Phoenix: Author.
WREB (2009b). *Dental hygiene examination security*. Phoenix: Author.
WREB (2009c). *Dental hygiene examination 2009 candidate guide*. Phoenix: Author.
WREB (2009d). *Dental hygiene examiner manual*. Phoenix: Author.
WREB (2009e). *Dental hygiene examiners self-assessment exercises*. Phoenix: Author.
WREB (2009f). *Dental hygiene new examiner criteria/procedures self assessment*. Phoenix: Author.
WREB (2009g). *Dental hygiene school coordinator checklist*. Phoenix: Author.
WREB (2009h). *Policy and procedures manual*. Phoenix: Author.
WREB (2009i). *Policy guide: Dental hygiene, anesthesia, restorative exams*. Phoenix: Author.
WREB (Winter/spring 2009). *Newsletter*. Phoenix, Author.
WREB (Fall 2009a). *Dental student newsletter*. Phoenix: Author.
WREB (Fall 2009b). *Newsletter*. Phoenix: Author.
WREB (Fall 2009c). *Dental hygiene newsletter*. Phoenix: Author.
WREB (June 4-6, 2009). *Dental hygiene committee meeting minutes*. Phoenix: Author.
WREB (July 13-14, 2009). *Dental hygiene examination review committee meeting.* Phoenix: Author.
WREB July 16, 2009). *WREB Board of Directors Meeting minutes*. Phoenix: Author.
WREB (August 8-9, 2009). *Computer simulation committee meeting*. Phoenix: Author
WREB (November 7-9). *Computer simulation committee*. Phoenix: Author
WREB (November 20-22, 2009). *Computer simulation committee meeting*. Phoenix: Author.
WREB (December 17-19, 2009). *Dental hygiene committee meeting*. Phoenix: Author.
WREB (2010a). *2010 dental hygiene examination candidate guide*. Phoenix: Author.
WREB (2010b). *Examiner standardization*. Phoenix: Author.
WREB (February 2010). *An evaluation of the Western Regional Examining Board's 2008 Dental and Dental Hygiene Examination Program*. Phoenix: Author.