

An Evaluation of the
Central Regional Dental Testing Services
National Dental Examination

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

January 7, 2011

Acknowledgments

The evaluation of any examination program is an exhaustive process that entails the analysis of many documents, the analysis of data, observations, and discussions with those involved in test development and validation. I would like to thank Lynn Ray for her consistent and invaluable assistance in this process. Without her help, this report would not have been possible.

Dr. Thomas M. Haladyna
Phoenix, Arizona
January 2011

Table of Contents

| | |
|---|----|
| Introduction | 1 |
| Part I: Purposes of this Evaluation | 2 |
| Part II: Description of the CRDTS <i>National Dental Examination</i> | 4 |
| Part III: Validity and Validation | 7 |
| Part IV: <i>Standards for Educational and Psychological Testing</i> | 10 |
| Part V: Legal Defensibility | 12 |
| Part VI: Validity Evidence Bearing on All Four Tests (II, III, IV, and V) | 13 |
| 1. Content-related Validity Evidence | 13 |
| 2. Item Quality | 16 |
| 3. Reliability | 18 |
| 4. Examination Administration | 20 |
| 5. Selection, Training, and Retention of Examiners and Scoring | 21 |
| 6. Comparability | 24 |
| 7. Standard Setting | 25 |
| 8. Score Reporting | 26 |
| 9. <i>Candidate Manual</i> and Rights of Test Takers | 27 |
| 10. Security | 28 |
| Part VII: Validity Evidence Bearing on Part II: Endodontics | 30 |
| Part VIII: Validity Evidence Bearing on Part III: Fixed Prosthodontics | 33 |
| Part IX: Validity Evidence Bearing on Part IV: Restorative Examination | 37 |
| Part X: Validity Evidence Bearing on Part V: Periodontal Examination | 39 |
| Part XI: Summative Evaluation and Recommendations | 43 |
| References | 44 |
| Appendix: Archive of Cited Documents Providing Validity Evidence | 46 |

INTRODUCTION

Testing agencies provide important information to states concerning the competence of candidates for licensure to practice a profession in that state's jurisdiction. These professions include dentistry, dental hygiene, accountancy, architecture, medicine, teaching, social work, law, and law enforcement among many other professions.

CRDTS has engaged this author to conduct an independent evaluation of its dental clinical performance test, the *National Dental Examination (NDE)*. This author has conducted such evaluations previously and has written extensively on the subject.

This evaluation report has seven parts:

- Part I addresses why this evaluation is being conducted.
- Part II describes the *NDE*.
- Part III discusses validity and the investigative process known as *validation*.
- Part IV identifies professional testing standards that apply to this test.
- Part V discusses legal defensibility of CRDTs test scores usage.
- Part VI reports the validity evidence bearing on all four tests of the *NDE*.
- Part VII reports validity evidence unique to the Endodontics test.
- Part VIII reports validity evidence unique to the Fixed Prosthodontics test.
- Part IX reports the validity evidence unique to the Periodontic test.
- Part X reports validity evidence unique to the Restorative test.
- Part XI is a summative evaluation of the *NDE*.

References are provided at the end of this report. The appendix shows documents reviewed for this evaluation. These documents comprise some validity evidence bearing on the summative evaluation of the *NDE*.

PART I: PURPOSE OF THIS EVALUATION

Organizations like CRDTS provide information concerning the validity of test score interpretations and uses to its clients, which are its state boards. Although CRDTS and other regional testing services design, develop, and administer these tests to candidates for licensure, state boards have the responsibility for licensing these candidates. Thus, responsibility for validity is vested in state boards not the testing agency that produces and reports test scores. However, responsible testing agencies, like CRDTS, regularly engage in self-evaluation and external evaluation to assess its strengths and weaknesses and improve its testing program and by that validity.

Why has CRDTS engaged this evaluator to conduct an independent evaluation of the *NDE* Program? Periodic evaluation of any testing program is an effective way to gain insight into the quality of the program and discover ways to improve it. That is why testing experts have long recommended and endorsed the external evaluation of testing programs (Buckendahl & Plake, 2006; Downing & Haladyna, 1996; Madaus, 1992). Evaluation entails considerable study of documents, some data analysis, and discussion with representatives of CRDTS. In this report, many references are made to the testing literature that bears on this evaluation. Also cited are documents originating from both CRDTS and the test development organization known as the American Board of Dental Examiners (ADEX), with which CRDTS was formerly associated. Also data from CRDTS tests administered in 2010 were analyzed and are reported here.

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

1. Defining of a profession as to competencies needed to practice safely,
2. Development of an examination that validly measures competence, and
3. Validation of the interpretation and uses of examination scores.

One might think of validation as an investigation bearing on validity. Because no examination or battery of tests is completely valid for measuring professional competence and because no system of making pass/fail decisions is infallible, validation serves two very useful purposes: (1) It determines 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.

The main purpose of any testing agency for licensure like CRDTS is to increase the likelihood that the professionally licensed person will safely treat the public they serve. The content of a professional licensure examination is an ability–professional competence. Using the language of modern validity theory, professional competence in dentistry is defined in terms of how well the candidate can perform a representative set of tasks from a domain of tasks identified as critical to the profession and representative of dentistry. Each task requires the complex use of knowledge and skills. CRDTS has identified these tasks. The *NDE* is claimed by CRDTS to produce validly interpreted and used test scores for helping its member states make accurate decisions regarding the ability of candidates to practice dentistry in their state.

Part II: Description of the *National Dental Examination (NDE)*

To be licensed to practice dentistry in any state or United States' jurisdiction, a candidate has to meet many qualifications, which include passing a series of tests. The *National Dental Board Examination–NDBE* (Parts I and II) consists of two of these tests. Then candidates are also expected to pass a clinical performance test, which is developed and administered by a regional testing agency. CRDTS is a testing agency that is responsible for a clinical performance test in dentistry. CRDTS was established in 1972. As stated in its bylaws, state boards for dental licensing are its members. Its members meet annually in August (CRDTS, 2008).

The ADEX is an umbrella organization formed to design a national clinical dental examination. Evidence of the origin of the examination and its organization, structure, staff, and committees is presented in annual reports (ADEX, 2006, 2007, April 5, 2005; April 10, 2006; April 12, 2007; April 17, 2007; December 5-6-7, 2007; January 19, 2008a; January 19, 2008b; August 21, 2008). As of June 30, 2009, CRDTs severed its association with ADEX but retained much of the examination, in which they had actively participated during its development over a 4-year period. Up to that point, documentation of validity was done by ADEX. After that point, the responsibility for subsequent documentation and any modifications of the examination has been the responsibility of CRDTS.

The *NDE* is used to measure a candidate's clinical competence in dentistry. With the permission of candidates, scores are sent to appropriate member states and other participating states. These states use this information to make pass/fail decisions about licensing each candidate. More information about CRDTS can be found on its website: <http://www.crdts.org/>

The *NDE* consists of four clinical performance tests. Each candidate can achieve a score as high as 100 points on each test. The successful candidate is required to pass each test to qualify for licensure, and 75 is the cut score for making pass/fail decisions. In this section, the *NDE* is described. More detailed information about the examination can be found in the 2010 *Dental Candidate Manual* (CRDTS, 2010). A technical report also provides detail about the examinations (Klein, November 2, 2010). Issues are presented in this part of the report bearing on the characteristics of this testing program.

Origin of Current Examination

The current examination is based on the collaboration of members of ADEX, which includes CRDTS. References in this report to ADEX are made to cite evidence in various aspects of the test development, administration, and scoring that occurred before June 30, 2009, when CRDTS severed its ties to ADEX. One report by ADEX (January 10, 2008) provides an example of examination review and recommendations made to ADEX that bear on the current examination. ADEX has conducted many meetings for the purpose of designing and improving the testing program (ADEX, 2006, 2007, April 5, 2005; April 10, 2006; April 12, 2007; April 17, 2007; December 5-6-7, 2007; January 19, 2008a; January 19, 2008b; August 21, 2008).

Traditional versus Curriculum Integrated Formats

All candidates for licensure who take the *NDE* have the option of taking the traditional format or the Curriculum Integrated Format. The traditional format requires that candidates take all examinations at the end of their dental education. The Curriculum Integrated Format allows earlier administration of examinations for qualified candidates who are currently enrolled in dental school with the caveat that if a candidate fails, there is remediation and retesting.

Conjunctive Versus Compensatory Scoring

For any test, the test agency can require candidates for licensure to pass each section of a test. This requirement is known as *conjunctive scoring*. Conjunctive scoring is a high standard, because poor performance in any section is not tolerated and results in failure. The rationale for conjunctive scoring is that low performance in any section may lead to unsafe professional practice. Pros and cons of conjunctive scoring are numerous (see Haladyna and Hess, 1999). The purpose of a licensing examination is to screen candidates who may practice unsafely and harm patients. Conjunctive scoring is justified, if in the judgment of CRDTS Board, low performance in any of the four tests signals potential unsafe professional practice. However, achieving high reliability for each test in the battery of tests is challenging.

Compensatory scoring requires that a pass/fail decision be made on the total score. Variable performances in sections of the test are considered irrelevant. Low performance in one area can be made up by higher performance in another area. All the candidate has to do is to earn a total score high enough to meet or exceed a single cut score. When compared with conjunctive scoring, compensatory scoring leads to a higher percentage of passing scores. Compensatory scoring is easier to do, and it is less costly than conjunctive scoring. Conjunctive scoring is more demanding of resources and test development. Compensatory scoring is also more reliable than conjunctive scoring because the results of each test are combined into a single test score.

CRDTS has determined that a conjunctive scoring model be used. The rationale for such an important decision follows a line of reasoning that asserts that low performance in any of the four areas is not satisfactory. Patient health and safety are jeopardized if performance is low in any one of the four test areas. State boards have the ultimate responsibility for deciding who passes and fails. They alone decide whether the use of conjunctive or compensatory scoring model is appropriate to their needs. State boards also determine their cut scores.

The Rating Scale Used for Scoring Candidate Performance

CRDTS has established a four-point rating scale to evaluate candidate performance on many tasks (test items) (*2010 Dental Candidate Manual*, p. 7). The scale is summarized briefly here but this reference provides more details about this rating scale.

- 4—Satisfactory.
- 3—Minimally acceptable
- 1 -Marginally substandard
- 0 -Critical deficiency

If the three examiners agree, that score is assigned. If two of the three examiners agree, that score is assigned. If the three examiners disagree, the median is assigned. If two or three examiners independently agree on a rating of a critical deficiency, the score is voided for that procedure and for the entire examination section.

Part I: *National Dental Board Examination (NDBE)–Parts I and II*

CRDTS is not involved in the development, administration, scoring, or validation of either of these two tests. As noted previously passing Parts I and II are initial steps in the licensing process. More information about this examination program can be obtained from the following website: <http://ada.org/110.aspx>. CRDTS recognizes that diagnosis and treatment planning skills are a critically important component of competent dental practice. It is CRDTS' conviction that the *NDBE* Parts I and II are the best assessment of these skills. While CRDTS does not require successful completion of *NDBE* Part I and II as a prerequisite for taking CRDTS' clinical examination, that requirement is maintained by all CRDTS' member and participating state boards as a prerequisite for licensure eligibility

Part II: Endodontic Examination of CRDTS' *NDE* (100 points)

This mannikin-based test consists of two activities (subtests) described in the *2010 Dental Candidate Manual* (CRDTS, 2010b): anterior endodontics and posterior endodontics. The anterior endodontics subtest is scored in terms of 12 criteria using the four-point rating scale. The second subtest (posterior) is scored using four criteria also using the four-point rating scale. As just noted, three well-trained and validated examiners provide ratings.

Part III: Fixed Prosthodontics Examination of CRDTS' *NDE* (100 points)

This mannikin-based test consists of three procedures (cast gold crown, porcelain-fused-to-metal crown preparation, and ceramic crown preparation). For the first procedure 11 criteria are used. For the second and third procedure 10 criteria are used respectively. The same four-point rating scale is used.

Part IV: Periodontal Examination of CRDTS NDE (100 points).

This patient-based test requires the candidate to do an extra/intraoral assessment (10 points), periodontal measurements (16 points), scaling (56 points), plaque removal (6 points), and tissue management (12 points).

Part V: Restorative Examination of CRDTS NDE (100 points).

This patient-based test requires the candidate to complete four procedures as noted in the *Dental Candidate Manual* (CRDTS, 2010b, p. 8). The score for each set of criteria is the ratio of points earned/points possible. The total score is computed by totaling all points earned across the four procedures and dividing by the total possible points.

Penalty Deductions

CRDTS has developed a system of scoring where deductions are made where performance is deficient (CRDTS, 2010, pp. 10-11). An extensive system of penalty points exists for the Restorative test. The *2010 Dental Examiner's Manual* (CRDTS, 2010, pp. 20-21) and the *2010 Chief Examiner's Manual* (CRDTS, 2010) also provides information about scoring and penalty points. The candidate's professional demeanor is also evaluated during the testing sessions. In each instance, examiner consensus must exist for penalty points to be assessed.

PART III: VALIDITY

The most important concern in any examination is *validity*. An examination score should accurately describe a candidate's level of competence as a dentist. The decision to pass or fail any candidate must be as valid as possible. Validity is a judgment concerning to what extent any test score or pass/fail decision is accurate. Therefore, the focus of this evaluation is validity. All other ideas about test quality are subsumed under validity, including reliability.

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 for validity were applied in this evaluation.

What does an examination score obtained from the *NDE* 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* or *test validity* is inappropriate. Validity focuses on the meaningfulness of an interpretation and the reasonableness of using the test score to pass or fail a candidate.

Validation

As noted previously, validation is an investigative process intended to facilitate an evaluation of validity. The first step in validation is to define dentistry. The standard for definition is the completion of a study known as a *practice analysis*—a survey of those in the profession whose judgments we value (Raymond & Neustel, 2006). To validate an interpretation of a test score for a candidate, specific information is required:

1. an argument that lays out what competence is to be measured and how CRDTS plans to measure it;
2. a claim is made that CRDTS' *NDE* test scores are valid measures of the competence of a dental candidate, and that it is defensible to use a test score for making a pass/fail decision;
3. validity evidence is assembled and 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. Validity studies are one remedy. By studying negative evidence and seeking remedies, validity is increased.

Table 1 shows the constituent elements in validation.

| Table 1: Validation of CRDTS's <i>NDE</i> | |
|---|---|
| Argument | The American Dental Association administers a <i>National Board Dental Examination</i> . This examination measures the knowledge and skills thought to be necessary for safe and competent dental practice. This examination derives principally from a practice analysis of the profession of dentists. The CRDTS <i>NDE</i> is a clinical performance examination intended to measure dental clinical competence directly. These two examinations represent complementary aspects of dental competence. CRDTS's <i>NDE</i> is the capstone in this licensing process for licensed dentists. |
| Claim About Validity | CRDTS claims that candidate scores from its <i>NDE</i> represent dental clinical competence. The results of the test can be used with confidence by participating states, along with other criteria, to make licensing decisions for candidates. |
| Evidence Supporting the Argument | This evaluation report provides validity evidence of many types that are based on national test standards. CRDTS's 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, CRDTS should consider threats to validity and act accordingly to diminish the threat. By that, CRDTS strengthens the evidence supporting the argument and the claim for validity. |
| Lack of Evidence | If evidence is missing, it is the responsibility of CRDTS to gather such evidence in the future as it contributes to increasing validity. |
| Summative Judgment | This evaluator considers the argument, claim, and evidence before making a judgement about validity of CRDTS scores as (1) a measure of clinical competence, and (2) for use by participating states in making pass/fail decisions. |

Validity Evidence Used in This Evaluation

To organize validity evidence, the following categories are presented: content, item quality, examiner consistency and reliability, examination administration, training of examiners and scoring, comparability, standard setting, reporting, candidate guide, and security. This body of evidence is evaluated holistically. 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 what 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. In any evaluation, the evaluator is responsible for truthfully exposing threats to validity (Cronbach, 1988). 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 tasks that comprise dental competence. This part of the evaluation seeks to uncover evidence that may undermine validity. Naturally, CRDTS 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 for validity. Every examination is only as strong as its weakest link.

Fidelity is a term we use to assess the connection of the tasks on the examination to the definition of competence for dentists. If we used a multiple-choice examination of scientific knowledge or a multiple-choice examination of professional knowledge, we would not be representing dental clinical competence adequately. These multiple-choice tests under-represent the construct of dental competence. That is why the CRDTS *NDE* is a necessary licensing requirement. When we combine the results of the *NBDE* with *NDE*, we have important complementary pieces of information that provide adequate representation of the construct of dental competence. Thus, participating states use both the *NBDE* and the *NDE* examinations due to their complementary nature.

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 2 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 CRDTS used for the validity claim stated in this evaluation. This chapter has many categories of validity evidence. CRDTS's annual technical report is one source of documentation. This report is another source. CRDTS 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 CRDTS'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 dentists, the issue of fairness is an important one. The design and administration of the *NDE* 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 *NDE*.

Chapter 9: Linguistic background. As this clinical performance examination involves patient treatment under simulated natural conditions involving patient-dental 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 a dental school. Foreign trained candidates often have difficulty with the English language. These candidates should also be treated fairly. While most candidates seeking licensure in the United States have adequate skill in the English language, having linguistic or hearing difficulties sufficient to require an interpreter is not uncommon for patients. CRDTS permits interpreters to be available as needed outside the candidate's operatory or the examiner station. 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 dentistry.

| Table 2: 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 Candidate Manual* (CRDTS, 2010b) 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 CRDTS'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 CRDTS'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 CRDTS's responsibility. However, CRDTS is responsible for providing all participating states with such information that supports their uses of examination scores. CRDTS's *Dental Candidate Manual* (CRDTS, 2010b) is published every year and provides much information. CRDTS's website also provides public access to documents.

PART V: LEGAL DEFENSIBILITY

Besides providing the highest quality examination possible, CRDTS 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 program.

Validation is an effort to provide evidence that supports the examination and its purpose. By undertaking a validation, CRDTS 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 validity evidence is available that supports validity, 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, CRDTS very effectively reduces the threat of legal action. Mehrens and Popham (1992) provide a useful discussion of legal threats and validity. CRDTS has made public its validity evidence in technical reports and evaluations such as this one. CRDTS's website is very informative (See <http://www.CRDTS.org/>).

PART VI: VALIDITY EVIDENCE BEARING ON ALL FOUR TESTS

Because the *NDE* consists of four tests, the body of validity evidence is organized in the following way. In this section, all evidence bearing collectively on all four tests is reported. Part VII covers validity evidence unique to the Endodontics test. Part VIII deals with evidence unique to the Fixed Prosthodontics test. Part IX treats the Periodontics test. Part X addresses the Restorative test.

As noted previously, evidence should be considered holistically in the judgment of validity. However, threats to validity may be serious and should be considered *weak links*. The summative judgment offered at the end of this report is based on the evaluator's judgment of all evidence with respect to the claim for validity. For the purpose of offering 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 clinical examination should identify a domain of tasks performed by a competent dentist. This domain is known as the *target domain*. Ideally, the tasks in the domain are organized by important content topic descriptors. These tasks are prioritized according to relevance to the profession and how frequently the tasks are performed in regular professional practice. 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 CRDTS'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 CRDTS expected to define clinical competence in dentistry, 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.

ADEX is the organization responsible for the design of this examination, which is currently used by CRDTS. In their meeting minutes, evidence bearing on all types of validity evidence is present (ADEX, 2006, 2007, April 10, 2006; June 23, 2006, August 26, 2006; April 12, 2007; December 5-7, 2007; January 19, 2008a; January 19, 2008b; January 22, 2008; August 21, 2008). Their examination review committee in particular provides a long list of modifications in the examination bearing on content, administration, examiner qualifications, examiner training and calibration, scoring, and quality assurance.

Practice Analysis

Klein (April 15, 2008, November 2, 2010, pp. 28-34) reported that a practice analysis was conducted by the Buros Institute. The result of this analysis was used to develop the test items (tasks) on the current four examinations. This survey is reported to have been conducted in four steps. First, subject-matter experts (SMEs) were consulted to generate a list of entry-level judgments, procedures, and skills required in dentistry. Second, a survey was designed based on the results of step one. Third, data was collected from a national sample using a representative sampling plan. Fourth, the results were summarized for designing the tasks on the test. ADEX meeting minutes provide evidence of a deliberate, planned analysis of results that articulate findings with the content of the examination (ADEX, January 19, 2008, January 22, 2008).

Structural Evidence

A major consideration in the design of any testing program is the theoretical and empirical structure of test data. Is dental clinical competence a single entity consisting of highly related tasks? Or is competence a family of relatively independent tasks, each of which is important in achieving a satisfactory level of competence? A study is reported here.

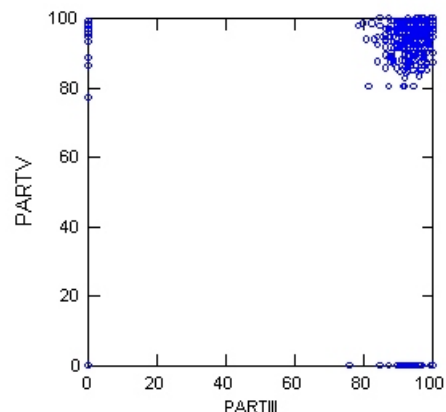
Table 3 provides descriptive statistics for the Part II, III, IV and V.

Table 3: Descriptive Statistics for the Four Tests of the NDE

| | Endodontics (II) | Prosthodontics (III) | Periodontics (IV) | Restorative (V) |
|-------------|------------------|----------------------|-------------------|-----------------|
| Candidates | 461 | 464 | 422 | 467 |
| Low Score | 0 | 0 | 0 | 0 |
| High Score | 100 | 100 | 100 | 100 |
| Mean | 88.7 | 89.2 | 94.7 | 86.1 |
| Stand. Dev. | 24.4 | 20.3 | 8.0 | 27.4 |
| Skewness | -3.2 | -4.0 | -4.8 | -2.8 |

As shown in the table, some candidates scored zero and failed that part of the *NDE*. Perfect scores are often achieved. The mean score for each test is well above the cut score (75). The variability of scores is primarily due to the preponderance of zeros for any test. For instance, for Endodontics, there were 31 zeros and then scores ranged from 68.8 to 100.0. The majority of scores for all tests were very high. Considering that the candidates have high ability and are well-trained, the expectation is that performance will be very high. Unfortunately, a small percentage fail each part, but are given the chance to remediate and re-take the test. These results show both remediated and final tests for some candidates. The skewness statistics shows negative skew—which means that most scores are very high.

Correlations among the four tests are close to zero. The scatterplot for the Fixed Prosthodontics and Restorative tests is very informative. The cluster on the upper right shows passing candidates on both parts. Some candidates failed Prosthodontics with zero and also passed restorative. Some candidates passed Fixed Prosthodontics with a score of 100 but failed the Restorative test. One candidate failed both parts (seen in the lower left side of the scatterplot). As zeros are the result of a special scoring rule, if we remove the zeros from the analysis, the upper right corner of the scatterplot shows the restriction in range that exists. Disregarding the zeros, note that very few scores fall below 75.



Conclusion

The practice analysis was conducted and was cited in a technical report (Klein, April 15, 2008; Klein, November 2, 2010). The second Klein report provides considerable detail on the practice analysis and its results. The study appears to meet standards for certification/licensure type examinations. To study the structure of the performance tasks in the four tests of this examination, descriptive statistics were computed. The results show four independent parts of dentistry. There is no evidence of relatedness between and among any of these four tests. Thus, the decision of CRDTS to use a conjunctive scoring model is supported by these results.

Recommendation

A new practice analysis is needed to update and confirm the judgment of CRDTS' Board that the tasks performed on CRDTS *NDE* are appropriate for the profession of dentistry. The results of this survey should be reported and published. A good place to archive the report is on CRDTS website. The Steering Committee and the three examination committees of the dental examination review committee should provide evidence that these results were considered in the current design of each of the four examination and any revisions that follow.

2. Item Quality

The kinds of test item formats used in any licensing examination can vary significantly (Albino, *et al.*, 2008). These formats include performance, multiple-choice, objective structured clinical examination, laboratory exercises, manikin, chart-stimulated evaluation, longitudinal, repeated observations, and portfolio to mention a few. No matter the specific formats employed, a rationale has been provided that show that each test item elicits the desired behavior for a specific task in the domain of relevant tasks defining the profession. The connection of each task on each of the four tests should connect to the practice analysis results.

Once item formats have been identified for any test, evidence bearing on item quality needs to be collected and organized. Items should undergo systematic development that depends on the expertise of CRDTS'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 knowledge, skills, and abilities 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 are field tested to assure its ability to discriminate between high- and low-performing candidates.
7. Most important, these items should have high fidelity with the criterion behavior intended—actual dental practice.

Fidelity

Tasks on any clinical test such as CRDTS should resemble those tasks performed by dentists in practice. If the tasks possess fidelity with criterion behavior, part of the validity argument is that the content of the *NDE* has high fidelity with the tasks performed by dentists in practice. A review of these tasks and prior committee activities supports the fidelity argument. The tasks performed on the examination are either identical to tasks performed on actual patients in dental practice, or, in the instance of manikin-based testing, have high fidelity with actual patient practice (in the judgment of the Board).

Examination Review Committee Meetings

CRDTS and its predecessor ADEX have had examination committee meetings where each of the four examinations are reviewed and modifications are proposed and made. (ADEX, April 10, 2006; June 23, 2006; August 26, 2006; April 12, 2007; April 17, 2007; December 5-7, 2007; January 19, 2008b; CRDTS, November 8-9, 2008; April 17-18, 2009; August 2009; January 16, 2010; April 19, 2010; August 26, 2010; October 22, 2010). Evidence is provided in

each of these reports of the outcomes of each subcommittee. The conclusions and recommendations bear on many aspects of test development including item quality and how the examination is administered and scored.

Weighting of Test Items

This topic is very important as weights assigned to items have consequences for candidates. In the development of this examination, ADEX and CRDTS has carried out evaluations of different weighting systems and arrived at the present one (ADEX, April 5, 2005, CRDTS, April 12, 2005). Since the original examination was developed by ADEX, CRDTS has reviewed and revised the original weighting of test items. There is evidence of a deliberate process of review, evaluation, and revision (Ray, March 27, 2008). The weighting of any test item is a matter of professional judgment by subject-matter experts. The decisions for the current weights for test items are the result of a deliberate process by the examination review committee during their frequent meetings.

Conclusion

Item development appears to have been adequate. The items were based on an earlier practice analysis. CRDTS has administered this examination for many years, and its association with ADEX has served to improve the examination process and sharpen the effectiveness of its items for measuring important clinical skills. Greater and more comprehensive documentation of item development and validation is very desirable. Regularly published meeting of the examination review committee and its subcommittees is the most dependable and useful way to document item quality evidence. An annual technical report should also report work on item quality.

3. Reliability

Every test score has an unknown degree of random error and a true score. This error can be positive or negative and large or small. There is no way to discover how much error in a test score or the true score (unless every task in the target domain is administered and perfectly scored). For a candidate whose test score is close to 75 (the cut score), we have a concern that a pass/fail decision might be incorrect due to random error. We have two kinds of errors of classification for pass/fail decisions. Either the passing candidate receives a fail decision when the true score is passing (equal or above 75) or the candidate receives a passing decision when the true score is failing (below 75). We call these classification errors Type 1 and Type 2. Reliability affords us some understanding of the risk of misclassifying candidates whose true scores are close to the cut score. For the other candidates, their scores are sufficiently high or low enough where there is little risk of misclassification regarding passing or failing.

For a complex performance examination as the *NDE* that has four distinct tests that require pass/fail decisions for each candidate, the risk of misclassification is greater. Fortunately, CRDTS has taken steps to ensure that reliability is high and the risk of misclassification is reduced.

1. CRDTS uses three examiners for each observation. This step ensures a high degree of internal consistency in ratings that is crucial in establishing reliability. Results of examiner consistency are reported in appropriate sections of this report for each of the four tests.
2. CRDTS has many observations (test items) per test. Reliability benefits by having many observations.
3. CRDTS has special scoring rules for critical deficiencies. This scoring rule results in automatic failure if two or three examiners agree that a performance justifies a rating of zero—indicating a critical deficiency (*CRDTS (2010) Candidate Manual*, p. 8).

The structure of the data for each of the four tests is not internally consistent, so conventional reliability estimates are not useful. Conventional reliability estimation depends on internal consistency of item responses. That is to say, item responses tend to be highly intercorrelated. Instead a more appropriate technique for estimating reliability is stratified alpha (Haertel, 2006, pp. 76-78). Haertel asserts that conventional reliability methods greatly underestimate reliability; whereas stratified alpha does not.

As the candidate pool consists of very high-performing candidates, test data is very heavily negatively skewed. Most statistical techniques (involving reliability and correlation) depend on a normal distribution with considerable variation of test scores. CRDTS test scores are very restricted. Thus, reliability estimates tend to be very low because of a lack of internal consistency and the skewness in scores.

Reliability is not an end; it is a means to an end. The objective of estimating reliability is to obtain an estimate of the margin of error around the cut score so that the Board can assess the

risk for misclassifying candidates whose true scores are close to the cut score of 75. Once reliability is properly estimated, the degree of random error is estimated and used to study the number of candidates whose observed scores falls near the cut score. Hopefully, the margin of error is very low and the number of candidates whose scores fall into this margin near the cut score is small.

Remedy

To estimate reliability of test scores appropriately the following steps were taken with the following justifications:

1. Zeros were removed from analyses as these were established by an independent scoring rule that has no bearing on true and random error scores.
2. Internal consistency reliability (alpha) was estimated for each trio of examiner ratings as these ratings are supposed to be internally consistent. Evidence will show that this is true.
3. Stratified alpha was used to estimate reliability for each subtest. These coefficients tend to be higher than a conventional coefficient alpha.
4. Stratified alpha was used with the total score (sum of subtests).

Cautions

The methods used to estimate reliability are extremely complex and make great demands on the data. The results reported in subsequent sections for each test provide insight into the reliability bearing on candidate scores that roughly range from 65% to 100%. Lower scores are not useful or relevant to the results presented due to the fact that these special scoring rules were used. These lower scores have resulted in failing decisions.

Results

Results are reported in appropriate sections of this report. In these sections, stratified alpha is very high despite the heavily skewed test score data. The margin of error is small, and the number of candidates observed close to the cut score are few.

Conclusion

The frequency of observations by examiners and the use of well-trained examiners to achieve consistency in ratings makes for highly reliable subtest scores and highly reliable test scores. This result in turn makes the standard error of measurement around the cut score minimal. Few candidates are in jeopardy of being misclassified. Thus, reliability appears to be a very strong aspect of this testing program. However, because the scoring system is so complex, estimating reliability is problematic. The results reported are very impressive given the complexity of each of the four tests, the random error due to human scoring, and highly skewed performances of candidates.

4. Examination Administration

This standardized examination has been administered over many years. During that time, the examination administration has been improved. When ADEX was responsible for the examination, regular meetings of various committees contributed to improving examination administration (ADEX, April 10, 2006; August 26, 2006; April 12, 2007; April 17, 2007; December 5-7, 2007; January 19, 2008a; January 19, 2008b; January 22, 2008; August 21, 2008). When CRDTS severed its ties with ADEX, a new Examination Review Committee and its subcommittees met regularly to improve examination administration (CRDTS, November 8-9, 2008; August 2009; April 17-18, 2009; January 16, 2010; April 29, 2010; August 26, 2010; October 22, 2010).

Another useful source of information about administration is the *2010 Dental Examiner's Manual*. This 98-page booklet provides background information about the examination, administration policies, examiner criteria, examiner responsibilities, among many other details of examination administration. The booklet also deals with mannikin and patient-based procedures and each of the four tests in this examination program.

Another useful document is the *2010 Chief Examiner's Manual*. This notebook contains more than 100 pages of information about the role of the chief examiner from three months prior to the examination to after the examination. The responsibilities are considerable. The notebook provides enormous detail and support for examination administration. Forms, instruction, guidelines, and criteria are included and organized by tabs. Finally, the *Dental Examiner Manual* (CRDTS, 2010) provides a wealth of information about examination administration.

Conclusion

The examination administration is very well organized. This is a mature testing program that has reached a high level of proficiency in examination administration. No recommendations are offered here.

5. Selection, Training, and Retention of Examiners and Scoring

This section addresses how examiners are identified and chosen, how they are trained, how they are evaluated, and then retained or dismissed, and issues of scoring. These activities are crucial to validity. According to the *Examiner's Manual* (CRDTS, 2010c, p. 11), CRDTS examiners are required to calibrate prior to every examination, no matter how recently they may have examined. Examiners are given specific team assignments for any given examination. Those assignments are to serve as a Restorative Examiner, Periodontal Examiner, or Clinic Floor Examiner. Usually, examiner assignments remain consistent throughout the year, unless circumstances require some adjustments (*2010 Examiner's Manual, Page 11*). The examiners go through general calibration together, led by the Chief Examiner. Then, they break into separate teams for four to six hours of calibration led by a Team Captain. The calibration exercises for each team are custom-designed PowerPoint presentations and a script and post-test is provided for the Team Captain. The calibration exercises are designed and updated annually by subcommittees of the ERC.

Selection and Retention of Examiners

One study established the practice of using three examiners (Klein & Bolus, May 9, 2008). As results show, this recommendation is very sound. Reliability is greatly aided by having three examiners per observation.

As described in a technical report (Klein, November 2, 2010), CRDTS and other regional testing agencies have written criteria for examiner selection and retention. These procedures are consistent with the American Association of Dental Examiners (AADE) published guidelines (AADE, 2005). Eight criteria are used to qualify examiners. Examiners must agree to conditions of engagement and follow strict protocols for training and retention.

Training and Evaluation of Examiners

CRDTS has an extensive system for training and evaluating examiners. Each examiner receives a copy of the most current *Dental Examiner's Manual* (CRDTS, 2010). Also, new examiners are given a Power Point presentation (CRDTS, 2010). This system includes home study and onsite opportunities for examiners. Examiner calibration is a fundamental part of this training. Examiners are trained to examine accurately and consistently with other examiners. All examiners receive detailed diagnostic feedback on their performance. All examiners participate in team meetings and work as teams. At the end of each examination year, all examiners are given a profile of their performance. This information is used for remediation of examiners and also for retention.

To augment this training, CRDTS has several annual publications that provide feedback to examiners on the extent to which they overrate or underrate candidates (Ray, 2010a, 2010b). Rater errors are a major threat to validity, and CRDTS's efforts to eliminate or reduce rater errors is laudatory.

Scoring

Extensive committee work was reported on all aspects of scoring (CRDTS, April 17-18, 2009). The *Dental Candidate Manual* (CRDTS, 2010b) provides the most recent update of the conditions for scoring including examiner ratings and penalty point assessments. All these decisions were reached by committee consensus and then approved by the Board.

Scoring is done on site and ratings are recorded electronically. According to Klein (November 2, 2010), after every examination there is verification and post examination review. All scores are rechecked. This effort seeks to uncover irregularities or errors in computing a candidate's score. All failing scores are subjected to manual verification by professional dental personnel.

Quality Control

All candidates are subjected to a multi-step process for standardization and calibration that is designed to produce accurate and consistent ratings of candidate performance. Exercises are designed and used during a two-day orientation of Chief Examiners and Team Captains. Chief Examiners contribute to the development of these exercises. Each year the exercises are reviewed, evaluated, and revised-if necessary. At the same time the *Dental Candidate Manual* is also revised (CRDTS, 2010b).

After the examinations are administered, CRDTS annually produces reports of examiner performance, which are intended for examiner self-assessment (Ray & Cobb, 2007). These results are also used to evaluate examiners and to inform decision-making for future examiner assignments. Results of the 2007 study show a range of examiner performance from as low as zero disagreement with as high as 16.5% disagreement. A set of reports for 2010 show similar findings (Ray, 2010a, 2010b). A very useful feature of these reports is the presentation of graphs showing degrees of leniency and severity in examiner judging. Such information can be very useful in refining training and improving examiner consistency or, if justified, removing examiners who are in consistent. Such reports are very useful for quality control.

CRDTS maintains an Examiner Evaluation and Assignment Committee (EEAC) that meets annually to review examiner profile reports, with additional meetings as needed to assign examiner teams for every test site. The EEAC reviews every examiner's individual profile, makes decisions regarding their effectiveness, looks for emerging leadership qualities as Team Captains or Chief Examiners, and assigns them as a Periodontal, Restorative or Clinic Floor Examiner based on their interests, skills and experience so that balanced teams are assigned to every site. They also review each examiner's Peer Evaluations, which are part of the profile reports. Every examiner is asked to evaluate their fellow team members at the close of each exam. These Peer Evaluations focus on the examiner's behavior, preparedness, adherence to protocol, and work ethic. The EEAC is empowered to change an examiner's assignment if they are not functioning well in a particular role, they may send letters to those examiners who are outliers in their profile reports, or terminate the examiner's assignments if their results or behavior is not appropriate.

As stated previously, CRDTS has criteria for retaining examiners. Thus, examiners who fail to rate accurately and consistently are unlikely to be reappointed.

Conclusion

The training of examiners by CRDTS is a highly refined activity that has received considerable attention over many years. The *2010 Dental Examiner's Manual* is an annual publication updated each year. It contains 52 pages of information and is very well organized. This document is supplemented with other materials used during training. Thus, training of examiners and their scoring is very effective. No recommendations for improvement are offered.

6. Scaling & Comparability

The validity of interpreting test scores is strongly dependent on having a test score scale that can be constant from one examination administration to another. With multiple-choice tests, it is customary to have multiple forms that have to be equated so the scale is constant from test to test and occasion to occasion. With a performance test, scaling and comparability of results are very different.

With the *NDE*, the same test items are used in each administration. The only variable is the examiner. All examiners come from a common pool of examiners. All are highly qualified and extensively trained. Their ratings are calibrated before they examine. CRDTS has checks and balances for examiners, and a feedback system to examiners alerts them to instances of leniency or severity in rating and inconsistency. Although the scoring system is complex, there is evidence of high examiner consistency and high reliability. Given all these actions to regulate and standardize the examination, the test score scale for Parts II, III, IV, and V seem comparable from test site to test site and occasion to occasion.

Fluctuations in the passing rates from site to site and from year to year may be explained by many factors including demographic characteristics of candidates, quality of their training, and minor variation in the quality of candidates taking training. Generally, stability in passing rates from year-to-year suggests the scaling is uniform. Dental schools would like to see a positive trend in passing rates as a function of their programs.

Conclusion

Scaling for comparability appears adequate given that this is a performance examination where the tasks are well known and frequently practiced by candidates. The use of three examiners helps stabilize the scale as ratings are highly consistent. Examiners are well trained and calibrated. All tasks are standardized. Although scoring is very complex, it too is standardized. The test score scales for each part are the same, as is the cut score. No recommendations are offered.

7. Standard Setting

States mandate passing scores of 75 on a 100-point scale as a legislative action for many testing programs under its aegis. Testing agencies develop procedures for the design of rating scales and scoring procedures that produce scores on a 100-point scale where 75 represents a very low performance. Thus, the argument is made that the cut score of 75 seems fair for determining levels of competency.

Conclusion

Because states mandate cut scores, testing agencies are placed in the uncomfortable position of having to justify the need for a passing score study that produces a result that very likely is not 75 on this 100-point scale. One remedy is to conduct a passing score study and slide the scale so that the cut score (say the recommended cut score of 72.4) becomes 75 on the new scale. That way, the testing agency has fulfilled their responsibility in performing a passing score study, and also, they report the practical cut score at 75.

CRDTS appears to be in compliance with respect to cut scores, but further work is needed to clarify the issue of a state's legislative mandate versus the testing standards espoused in the testing industry.

8. Score Reporting

CRDTS has three kinds of score reports:

1. one for dental schools, reporting scores of their current graduates;
2. another for state boards, reporting scores of all candidates at each test site; and
3. another for individual candidates. All appear to be designed for easy interpretation.

The dental school score report is very simple. It contains the candidate name, identification number and total score and subscores for each of the four parts. No summary statistics or normative information is provided.

The candidate score report has two versions: one for a candidate who passed and the other for the candidate who failed. Both reports are very simple and clear. Each report presents total score and subscores for each test, excepting the Periodontal test, which has no subscore information. At the bottom of the report is a comment section. For a failing candidate, justifications are provided about what performances led to a low score on a test or what caused the penalty deductions.

Conclusion

These score reports are excellent. No recommendations are offered.

9. Candidate Manual and Rights of Test Takers

The *Dental Candidate Manual* is the official publication for candidates (CRDTS, 2010b). This 120-page booklet contains many topics of interest and importance to candidates. Over 90 pages are devoted to the examination itself.

CRDTS website (<http://www.crdts.org/>) is also a source of information for candidates. The dental section offers candidates information about application and eligibility, the calendar for examinations, examination content, how the examination is scored, forms and manuals, online application, and orientation. The orientation consists of a Power Point presentation. Relevant forms are provided online. Under frequently asked question, the appeals process is described, and a summary description of this process is also provided in the *Dental Candidate Manual*. CRDTS has a review petition/appeals process for failing candidates who want to inquire about the accuracy of scoring. CRDTS will not rescore the examination, but will consider any evidence that points to alternative results.

Failing Candidates

As described previously, failing scores are verified. A candidate who fails any examination receives a report itemizing deficient performances. Applicants may question a failing score using the formal procedures that CRDTS has established and described in the *Dental Candidate Manual* (CRDTS, 2010b).

Conclusion

CRDTS provides its candidates a very well-prepared manual that provides a wealth of information about the examination schedule, process, and scoring. The website is very informative and appropriately provides many services online. The orientation appears intended for oral presentation during a meeting. Candidates have an appeal process that is clearly described. Overall, CRDTS does an excellent job of honoring rights of candidates for licensure.

10. Security

CRDTS has taken many steps to ensure security in examination development, administration, scoring, and reporting (CRDTS, 2010e).

CRDTS Central Office is in Topeka, Kansas. CRDTS office is located on the lower floor of a two-story building with a rear-entry access for pickups and deliveries. There are three satellite offices staffed by one staff member each. At least one full time employee is in the office during weekdays. The office is often locked. Visitors to this office can be observed before entering the reception area. There is a workroom for storage of examination materials, and there is secure off-site storage facility.

Staff members communicate by phone or via CRDTS web servers, which has a password protection for the transmission of confidential documents. Staff members also meet with CRDTS officials and visitors under supervision and attendance of staff.

Candidates must apply online. They must submit a notarized signature, two photographs, examination fee, and documentation of their eligibility. Candidates must view an orientation on line.

When candidates check in for the examination, they must present a photo identification from a government agency and an affidavit crediting the online orientation. Their identification card and photo are cross checked with the candidate list.

“Examination materials, such as Progress Forms and Flow Sheets, that are part of the candidate’s permanent record, are pre-printed with each candidate’s individual sequential ID number and a 10-digit computer ID number that is a secure coded version of their social security number. In addition, the electronic equipment for scoring the exam is pre-loaded with each candidate’s ID numbers, and the examiner ID numbers and names for all examiners assigned to the test site. This is done to ensure that all exam results are correctly identified” (CRDTS, 2010e).

CRDTS has metal trunks with combination locks for shipment of material from its office to test sites. CRDTS has a company that insures shipping and maintains security in transmission of its materials. As a wireless scoring system is used, materials needed for this recording of scoring are also packed and shipped in a secure way. CRDTS uses its own wireless network for transmission of data. The transmissions are constantly observed to ensure accuracy of data transmitted. All data are uploaded to a portable storage device and later uploaded into CRDTS secure scoring website before final scoring and reporting. CRDTS has back up systems for transmission and storage of data. Premier One Data Systems provides these services (<http://www.premier-one.com/>). All servers are protected by a variety of filters, spyware, and other defense systems to prevent unwanted intrusions. All documents are backed up.

Examination scores are processed and verified. All of this work is done on a secure website by staff, who have varying levels of password protection. Candidates have access to their scores using a password to access this information on the web. Scores are also sent to dental schools from CRDTS Archive and Document system. All of this information is stored in a separate filing system within the Archive and Document system. This separate system allows CRDTS to manage information for candidates and dental schools' interests without compromising internal security.

Candidates who try to bring prepared teeth to the examination in place of the teeth they are to prepare will be exposed because CRDTS's test modules have a special preparation applied to the model they use. For patient-based performance, all performance items are checked and recorded before the examination begins.

The examination materials may be lost or stolen in transmission. However, the only critical part of the examination is the electronically recorded results, which are transmitted electronically with ample backups and safeguards.

Acts of nature, such as hurricanes, tornadoes, or other disruptions happen. Although inconvenient, CRDTS has remedies for such events at no expense to candidates.

Conclusion

CRDTS is lauded for having an excellent system ensuring security in all phases of examination planning, development, administration, scoring, and reporting. No recommendations are made.

PART VII: VALIDITY EVIDENCE BEARING ON PART II: ENDODONTICS EXAMINATION

ADEX Examination Review Committee and Subcommittees met often. At these meetings, they reviewed and modified the Endodontics examination. They also provided evidence for content and item quality up to the time that CRDTS separated from ADEX (ADEX, April 5, 2005). Referring to the previous discussion of reliability in this report, this section reports on the consistency of examiners for the two subtests, the internal consistency of the examiners for the 16 criteria, the stratified-alpha reliability estimate for each subtest, and the stratified alpha reliability estimate for the total score.

Examiner Consistency–Subtest 1: Anterior Endodontics

Table 4 shows examiner consistency for each of the 12 observations. The consistency percentage is based on total agreement or one-point disagreement for which the median value is assigned. In other words, the following rating patterns are viewed as agreement: 444, 333, 111, 443, 334, 331, 113, 112. In each of these cases, the median value is used.

Table 4: Examiner Consistency for Anterior Endodontics

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|----|-------|-----------|------|------------|----------|
| 1 | 0.44 | 96.0% | 3.86 | 0.38 | -3.36 |
| 2 | 0.70 | 88.4% | 3.60 | 0.72 | -2.20 |
| 3 | 0.80 | 91.2% | 3.81 | 0.50 | -3.45 |
| 4 | 0.51 | 90.8% | 3.78 | 0.57 | -3.31 |
| 5 | 0.49 | 92.2% | 3.90 | 0.53 | -5.19 |
| 6 | 0.38 | 94.0% | 3.92 | 0.34 | -5.50 |
| 7 | 0.53 | 93.0% | 3.76 | 0.62 | -3.17 |
| 8 | 0.75 | 90.7% | 3.72 | 0.71 | -2.91 |
| 9 | 0.57 | 96.3% | 3.99 | 0.08 | -11.82 |
| 10 | 0.00 | 98.5% | 3.99 | 0.08 | -11.79 |
| 11 | 0.38 | 90.2% | 3.85 | 0.05 | -3.94 |
| 12 | 0.00 | 99.2% | 4.00 | 4.00 | 0.00 |

As shown above, the degree of examiner consistency is very high. Some variability exists. The lowest degrees of agreement (observation 2) is associated with the lowest candidate performances with the greatest variability. Although the alpha coefficients are moderate to low, the fact that we have 12 observations to estimate subtest reliability is very good.

Examiner Consistency Task 2

Table 5 presents the degree of agreement among examiners for the four observations on posterior endodontics. The degrees of agreement are lower when compared to anterior endodontics. Mean scores are also slightly lower. Alpha reliability estimates are slightly higher. This result comes because of lower agreement indexes and more variability in the scores on this subtest.

Table 5: Examiner Consistency for Posterior Endodontics

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|---|-------|-----------|------|------------|----------|
| 1 | 0.55 | 80.5% | 3.60 | 0.81 | -2.31 |
| 2 | 0.60 | 80.2% | 3.23 | 1.08 | -1.26 |
| 3 | 0.33 | 72.2% | 3.55 | 0.76 | -1.55 |
| 4 | 0.59 | 85.9% | 3.55 | 0.62 | -1.64 |

Reliability

The stratified reliability estimate for anterior endodontics is 0.742. The reliability estimate of the endodontics posterior subtest scores is 0.746. These coefficients are higher-than-expected given that the distributions of test scores for both subtests of the Endodontics test are very negatively skewed and scores are restricted in variability. These conditions attenuate reliability estimates.

As noted previously, stratified alpha is a suitable method for estimating reliability for two or more parts of an examination. Using the formula for stratified alpha, the total test score reliability estimate is 0.774.

Standard Error of Measurement

The margin of error surrounding a hypothetical true score (if we consider it to be the passing score of 75%) is what we might expect if a candidate has a true score of 75. That is to say, if a candidate has a true score of 75, by the unreliability of examiner judgments, what is the expected range of random error? The standard error of measurement was computed to be 7.17. Looking at the distribution of test scores for endodontics, 11 of 430 candidates have scores close to 75, which places them in this zone of uncertainty. These candidates have jeopardized their pass/fail status by dint of their performance. Improving reliability would reduce this standard error and, by that, result in fewer candidates being trapped in this zone of uncertainty. However, there is diminishing return from this strategy because so few candidates are close to the cut score of 75.

Conclusion

Examiner consistency is very high for the anterior endodontics and lower for the posterior endodontics. The subscore reliability estimates appear to be moderate, but this fact is mitigated by the fact that scores are very high skewed and restricted in variability. Total score reliability estimate is sufficient for making pass/fail decisions. No improvement is suggested for increasing reliability.

PART VIII VALIDITY EVIDENCE BEARING ON PART III: FIXED PROSTHODONTICS

Examiner Consistency–Cast Gold Crown

As with previously reported examiner consistency results, Table 6 shows that ratings are negatively skewed showing a preponderance of very high scores as expected from these candidates for licensure. The degrees of examiner agreement are very high, ranging from 81.3% to 100% with a median value of 93.0%.

Table 6: Examiner Consistency Cast Gold Crown–Subtest One

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|----|-------|-----------|------|------------|----------|
| 1 | 0.63 | 93.0% | 3.64 | 0.60 | -2.06 |
| 2 | 0.23 | 96.6% | 3.84 | 0.37 | -1.86 |
| 3 | 0.33 | 95.6% | 3.93 | 0.28 | -5.07 |
| 4 | 0.18 | 81.3% | 3.51 | 0.65 | -1.69 |
| 5 | 0.29 | 93.7% | 3.89 | 0.38 | -4.83 |
| 6 | 0.34 | 83.7% | 3.56 | 0.74 | -2.40 |
| 7 | 0.42 | 97.8% | 3.88 | 0.37 | -3.67 |
| 8 | 0.47 | 83.7% | 3.48 | 0.76 | -1.86 |
| 9 | 0.32 | 95.9% | 3.90 | 0.31 | -2.60 |
| 10 | 1.00 | 100.0% | 4.00 | 0.00 | 0.00 |
| 11 | 0.30 | 93.0% | 3.86 | 0.47 | -4.54 |

Examiner Consistency–Porcelain-Fused-to-Metal Crown Preparation

As with the previous subtest, examiner agreement for porcelain-fused-to-metal crown subtest is very high. Alpha coefficients are low due to the negatively skewed and restricted test scores. The most accurate indication of examiner consistency is the percentage agreement statistics because the candidate scores are so greatly skewed.

Table 7: Examiner Consistency for Porcelain-Fused-to-Metal Crown Preparation–Subtest Two

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|----|-------|-----------|------|------------|----------|
| 1 | 0.56 | 94.9% | 3.66 | 0.54 | -1.77 |
| 2 | 0.44 | 98.1% | 3.80 | 0.42 | -1.94 |
| 3 | 0.49 | 95.4% | 3.86 | 0.38 | -3.30 |
| 4 | 0.40 | 85.2% | 3.43 | 0.68 | -1.65 |
| 5 | 0.44 | 94.2% | 3.88 | 0.42 | -5.31 |
| 6 | 0.58 | 75.8% | 3.41 | 0.87 | -1.94 |
| 7 | 0.35 | 98.3% | 3.86 | 0.39 | -3.58 |
| 8 | 0.48 | 62.2% | 3.20 | 1.03 | -1.30 |
| 9 | 0.52 | 90.3% | 3.69 | 0.54 | -2.12 |
| 10 | 0.49 | 97.8% | 3.99 | 0.14 | -20.78 |

Examiner Consistency–Ceramic Crown Preparation

As shown in Table 8, the degree of examiner agreement was very high. Agreement percentages ranged from 84.2% to 98.8%. As noted previously, the lower percentages of agreement are related to slightly lower performances by candidates. For instance, observation 4 had a low mean (3.67) and a low examiner agreement (89.1). Nonetheless, these degrees of examiner consistency are very high. The same is true for observations 7 and 9.

Table 8: Examiner Consistency–Ceramic Crown Preparation–Subtest 3

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|---|-------|-----------|------|------------|----------|
| 1 | 0.41 | 94.4% | 3.56 | 0.57 | -1.37 |
| 2 | 0.31 | 96.5% | 3.76 | 0.43 | -1.23 |
| 3 | 0.50 | 98.8% | 3.93 | 0.28 | -4.82 |
| 4 | 0.34 | 89.1% | 3.67 | 0.59 | -2.23 |
| 5 | 0.44 | 97.6% | 3.83 | 0.52 | -3.87 |
| 6 | 0.53 | 94.7% | 3.84 | 0.47 | -3.76 |
| 7 | 0.48 | 84.2% | 3.72 | 0.68 | -2.91 |
| 8 | 0.44 | 90.8% | 3.77 | 0.53 | -2.96 |

| | | | | | |
|----|------|-------|------|------|-------|
| 9 | 0.52 | 85.2% | 3.78 | 0.64 | -3.40 |
| 10 | 0.62 | 90.1% | 3.88 | 0.58 | -4.76 |

Reliability

As noted previously, we have a complex structure for the data that includes two sets of linear combinations, a special case of reliability (See Haertel, 2006, pp. 76-78). To estimate reliability accurately, a two-stage stratified alpha was computed. At stage one, the stratified alpha was computed for each subtest. In stage two stratified alpha coefficient is shown in Table 9 below. As indicated there, the moderate reliability estimates for the three subtests contributed to forming a high stratified alpha coefficient for the total score. This is a very high reliability coefficient given the circumstances of this test: (1) scores are very negatively skewed, and (2) the range is very restricted. Also note, that zero scores were removed, because these scores are the result of a special scoring rule that does not bear on the reliability of test scores. The scoring rule involves two or three examiners independently deciding that a critical deficiency exists in the performance resulting in a score of zero. If a score of zero is assigned for one subtest, performance on the other two tests is irrelevant; the candidate fails that test.

Table 9: Sample Size, Mean, Standard deviation (stan. dev.) and Reliability Estimates

| Subtest | Sample Size | Mean | Stan. Dev | Strat. Alpha |
|--------------|-------------|------|-----------|--------------|
| 1. Cast Gold | 424 | 94.3 | 0.05 | 0.786 |
| 2. Porcelain | 432 | 91.9 | 0.06 | 0.738 |
| 3. Ceramic | 400 | 95.5 | 0.05 | 0.703 |
| Total Score | 462 | 93.6 | 0.04 | 0.855 |

The range of score for the Fixed Prosthodontics test was from 76.1% to 100%. The total score reliability of 0.855 this 33-observation subtest is high for a performance test despite the highly negatively skewed data.

Standard Error of Measurement

Given the cut score of 75 and the reliability estimate of 0.855 for the total score, the standard error of measurement is very small (1.5). Only two of the scores fall into that area of uncertainty.

Conclusion

Reliability for this examination is very high. Several factors contribute to this conclusion. First, there are three subtests. The more subtests, the higher reliability tends to be. Second, there are many observations per subtest, (33 for the first, 30 for the second, and 30 for the third). This makes a total of 93 observations using a four-point rating scale. Finally, instead of using

coefficient alpha that is known to underestimate reliability, stratified alpha was used which is more appropriate. The resulting reliability is very high despite two factors that are known to affect reliability negatively: (1) scores are very restricted due to the high performance, and (2) scores are negatively skewed. Most statistical indexes depend on variability and a normal distribution optimizes these statistics. In this instance, these limiting factors were overcome successfully.

Recommendation

No improvement in reliability is needed for this test.

PART IX: VALIDITY EVIDENCE BEARING ON PERIODONTICS

Examiner Consistency

This examination involves the detection of errors. Each candidate's performance is evaluated on 196 independent observations. Approximately 100% of these scorable performances achieved a maximum score. Thus, examiner consistency is nearly 100%. As there is no variation in these scores, examiner consistency indexes are not informative. Given that the tasks performed by these candidates is valid as supported by the practice analysis and the decision by CRDTS to use these performance tasks, rater consistency is as high as can be achieved.

Reliability

Given the constancy of the results, reliability cannot be estimated. However, if examiner consistency is near perfect, reliability must approach 1.00. In other words, there is little, if any, random error variance.

Standard Error of Measurement

Only 18 (4.1%) of the 434 candidates had scores in the five percentage point range around the cut score. Thus, the factor that places these candidates in jeopardy of being misclassified are penalty points. Candidates' scores at or near the cut score seem to result from penalty assessments.

Conclusion

Performance for most candidates is nearly perfect. Examiner consistency is nearly perfect. The only variation in test scores is due to penalty points. Penalties are assessed via a set of special scoring rules requiring examiner consensus. Given the high number of observations, the question arises: Is Periodontal test too long? Given the lack of variation in data and perfect scores, a very small sample of the current items (performance tasks) would provide an equivalent result. Perhaps the Board might consider a guideline that calls for all three examiners agreeing where penalty points are being assessed, as penalty points appear to have considerable influence on the pass/fail decision. Also, the Board might consider shortening this test as most performance items appear redundant.

PART IX: VALIDITY EVIDENCE BEARING ON PART IV: RESTORATIVE

Examiner Consistency–Amalgam Preparation and Finish

Table 10 presents an alpha internal consistency reliability estimate for each of the 12 observations. Also, the agreement indexes are presented. Note that these two statistics are complementary in nature. Alpha reliability estimates are low because means are extremely high and data is negatively skewed. In instances, where the means were lower (observations 3, 7, and 8), alpha estimates are higher and agreement indexes are lower. These results are explained by the fact that the lower means produces more variation in scores, which usually increases reliability estimates. Reliability coefficients strongly depend on score variation. In general, the result from Table 10 show a fairly high degree of examiner consistency.

Table 10: Amalgam Preparation Examiner Consistency

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|----|-------|-----------|------|------------|----------|
| 1 | 0.68 | 91.4% | 3.69 | 0.69 | -2.71 |
| 2 | 0.44 | 94.7% | 3.85 | 0.45 | -3.97 |
| 3 | 0.36 | 81.7% | 3.87 | 0.59 | -4.62 |
| 4 | 0.57 | 95.7% | 3.85 | 0.49 | -4.22 |
| 5 | 0.34 | 87.8% | 3.74 | 0.58 | -2.84 |
| 6 | 0.34 | 89.6% | 3.91 | 0.49 | -5.71 |
| 7 | 0.51 | 84.6% | 3.50 | 0.70 | -1.77 |
| 8 | 0.51 | 66.0% | 3.52 | 1.13 | -1.10 |
| 9 | 0.34 | 99.7% | 3.91 | 0.31 | -4.46 |
| 10 | 1.00 | 100.0% | 4.00 | 0.00 | 0.00 |
| 11 | 1.00 | 100.0% | 4.00 | 0.00 | 0.00 |
| 12 | 0.27 | 91.6% | 3.83 | 0.52 | -3.89 |

Table 11 reports similar results for amalgam finish. Where the mean test scores reach the ceiling of the scale (observation 3), reliability is not well estimated because there is very little variation in test scores. However, the agreement index is maximized. The results shown in that table again show a high degree of examiner consistency.

Table 11: Amalgam Finish Examiner Consistency

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|---|-------|-----------|------|------------|----------|
| 1 | 0.49 | 85.6% | 3.64 | 0.65 | -2.40 |
| 2 | 0.43 | 96.2% | 3.70 | 0.46 | -0.87 |
| 3 | 0.03 | 99.7% | 3.99 | 0.11 | -8.70 |
| 4 | 0.67 | 91.0% | 3.77 | 0.69 | -3.67 |
| 5 | 0.62 | 90.0% | 3.88 | 0.59 | -4.70 |

Examiner Consistency–Posterior Composite Preparation and Restoration

The second subtest has a similar pattern of examiner consistency for both parts, as shown in Tables 12 and 13. Examiner consistency appears to be very high.

Table 12: Posterior Composite Preparation Examiner Consistency

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|----|-------|-----------|------|------------|----------|
| 1 | 0.53 | 94.1% | 3.68 | 0.44 | -2.64 |
| 2 | 0.19 | 94.9% | 3.90 | 0.29 | -2.79 |
| 3 | 0.39 | 80.9% | 3.87 | 0.54 | -4.65 |
| 4 | 0.32 | 98.6% | 3.94 | 0.24 | -3.81 |
| 5 | 0.18 | 84.7% | 3.90 | 0.52 | -5.28 |
| 6 | 0.48 | 94.2% | 3.96 | 0.28 | -9.25 |
| 7 | 0.59 | 94.1% | 3.62 | 0.53 | -1.25 |
| 8 | 0.48 | 84.6% | 3.70 | 0.72 | -2.82 |
| 9 | 0.00 | 100.0% | 4.00 | 0.00 | 0.00 |
| 10 | 0.31 | 98.5% | 3.88 | 0.32 | -2.44 |
| 11 | 0.44 | 83.1% | 3.79 | 0.76 | -3.43 |

Table 13. Posterior Composite Restoration Examiner Consistency

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|---|-------|-----------|------|------------|----------|
| 1 | 0.32 | 78.5% | 3.47 | 0.76 | -1.78 |
| 2 | 0.60 | 96.1% | 3.83 | 0.44 | -3.09 |
| 3 | 0.21 | 97.7% | 3.96 | 0.19 | -4.90 |
| 4 | 0.00 | 100.0% | 4.00 | 0.00 | 0.00 |
| 5 | 0.32 | 79.0% | 3.70 | 0.70 | -2.82 |
| 6 | 0.59 | 83.0% | 3.75 | 0.83 | -3.05 |
| 7 | 0.65 | 93.8% | 3.62 | 0.65 | -2.17 |

Class III Composite Preparation

The third subtest has a similar pattern when compared to the two previous subtests. Examiner consistency tends to be very high where performance is high, and examiner consistency is lower where performance is lower. However, alpha is reasonably high when candidate performance is lower than the overall average. Tables 14 and 15 present the data for preparation and restoration of Class III Composite.

Table 14: Class III Composite Preparation

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|---|-------|-----------|------|------------|----------|
| 1 | 0.53 | 94.7% | 3.61 | 0.71 | -2.28 |
| 2 | 0.37 | 87.2% | 3.84 | 0.45 | -3.67 |
| 3 | 0.42 | 95.0% | 3.85 | 0.43 | -3.57 |
| 4 | 0.40 | 90.6% | 3.88 | 0.45 | -4.69 |
| 5 | 0.56 | 91.9% | 3.55 | 0.63 | -1.69 |
| 6 | 0.02 | 100.0% | 4.00 | 0.00 | 0.00 |
| 7 | 0.00 | 97.9% | 4.00 | 0.00 | 0.00 |

Table 15: Composite Finished Restoration

| | Alpha | Agreement | Mean | Stan. Dev. | Skewness |
|---|-------|-----------|------|------------|----------|
| 1 | 0.36 | 78.3% | 3.51 | 0.76 | -1.91 |
| 2 | 0.53 | 96.9% | 3.87 | 0.43 | -4.26 |
| 3 | 0.01 | 97.6% | 3.97 | 0.20 | -9.30 |
| 4 | 0.00 | 98.1% | 4.00 | 0.00 | 0.00 |
| 5 | 0.28 | 89.2% | 3.88 | 0.48 | -4.91 |
| 6 | 0.13 | 96.4% | 3.99 | 0.14 | -21.07 |

Reliability

The reliability of scores for this examination is the most complex of the four tests. First, for each subtest, a subtest score reliability estimate was computed using stratified alpha. Then, for each of the three combinations of subtests, reliability was estimated again using stratified alpha. Then a total score reliability estimate was done for each of the three combinations. Table 16 below shows the reliability estimates for the six subtests of the restorative examination.

Table 16: Reliability Estimates for Six Subtests of the Restorative Examination

| | Sample | Mean | Stan. Dev. | Reliability |
|---------------------------------|--------|------|------------|-------------|
| Class II Amalgam Preparation | 304 | 94.8 | 0.05 | 0.687 |
| Amalgam Finished Restoration | 298 | 94.9 | 0.07 | 0.697 |
| Posterior Composite Preparation | 128 | 96.3 | 0.05 | 0.640 |
| Posterior Composite Restoration | 131 | 94.0 | 0.07 | 0.709 |
| Class III Composite Preparation | 446 | 95.6 | 0.05 | 0.645 |
| Composite Finished Restoration | 438 | 96.9 | 0.05 | 0.481 |

Table 17 presents the reliability estimates for each subtest. As shown there high performance of candidates results in negatively skewed test scores that lower reliability estimates. Referring back to Table 16, the two components of Class III Composite (Preparation and Restoration) both have low reliability coefficients due to very high performance levels.

Table 17: Reliability Estimates for Total Restoration Scores

| Combination | Strat. Alpha | Mean | Stan. Dev. | Skewness |
|---------------------|--------------|------|------------|----------|
| Class II Amalgam | 0.769 | 94.5 | 0.05 | -1.47 |
| Class II Posterior | 0.812 | 95.0 | 0.05 | -1.98 |
| Class III Composite | 0.615 | 96.2 | 0.04 | -1.45 |

Because candidates have choices in the restoration examination, two reliability estimates exist. The first combination of subtests involves Class II Amalgam Preparation and Finished Restoration and Class III Composite Preparation and Finished Restoration. The stratified alpha coefficient is 0.895. The second combination of subtests involves posterior composite preparation and restoration. The stratified alpha is 0.747. The lower coefficient for the second combination is due to the high performance on both posterior subtests. The lower reliability estimates for class III composite preparation and finished restoration contributed to the overall lower reliability estimates of both combinations. However, reliability is attenuated by these heavily skewed distributions.

Standard Error of Measurement

Because of the two combinations and corresponding, we have two reliability estimates.

| Total Score Combination | Stratified Alpha | Standard Error | Candidates Detected |
|-------------------------|------------------|----------------|---------------------|
| First | 0.895 | .0648 | 2 |
| Second | 0.747 | .0677 | 1 |

Given the stratified alpha coefficients and the standard errors reported above, only three candidates had scores that were close to the cut score.

Conclusion

Examiner consistency is high, and reliability is sufficient for helping state boards make pass/fail decisions at the 75 cut score. This conclusion is justified in several ways. First, those who fail this examination fail by a large margin due to the critical deficiency condition. Second, most candidates score very high in these tests. Few candidates score near the cut score. Even if reliability were low, the risk of misclassification of candidates whose scores are close to pass/failing will be small. With the current design, content, item quality, administration, and scoring do not present problems. However, the complexity of this examination argues for a simpler test design where the estimation of reliability is more transparent to the stakeholders—the board, dental schools, and candidates. However, the current restoration examination does not present serious threats to validity in its current configuration.

VII: SUMMATIVE EVALUATION

CRDTS has designed and improved an examination that satisfies the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Moreover, the argument for validity presented in this report and the evidence assembled supports the validity of interpreting test scores as measures of dental competency and using test scores for pass/fail decision. CRDTS is lauded for achieving a high degree of validity with an examination program that has a long history with several transitions and milestones.

Some of the outstanding features of this program are as follows:

1. The decision to use a conjunctive scoring appears to be well grounded and defensible.
2. The number of observations for each test is very high and consequently supports reliability, which is very high.
3. Test design and development was done very carefully and is annually updated.
4. Tasks (test items) have high fidelity to the definition of professional dental competence.
5. Test administration runs very efficiently and effectively.
6. Selection and training of examiners has reached a very high level of excellence.
7. There is an effective system for retention and dismissal of examiners.
8. Scoring is very consistent and accurate.
9. Score reporting is simple and straightforward. Remedial/corrective information is provided to failing candidates.

Some caveats are offered here to sustain and improve the testing program: (1) Practice analysis needs to be conducted again. (2) The results of the practice analysis should be used to review the current examination performance items and verify their validity. (3) Technical documentation of all activities bearing on validity should be continued and expanded. All documents should be titled and dated. All meetings should have minutes and reports of meetings should be titled and dated. All documents should be archived. (4) Technical reporting should be done annually.

References

- Albino, J. E., Young, S. K., Neumann, L. M., Kramer, G. A., Andrieu, S. C., Henson, L., Horn, B., Hendricson, W. D. (2008). Assessing dental students' competence: Best practice recommendations in the performance assessment literature and investigation of current practices in predoctoral dental education. *Journal of Dental Education*, 72(12) 1405-1435.
- 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 .

- 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–AADE (2005). *Guidance for clinical licensure examinations in dentistry*. Chicago: Author.
- ADA (October 2004). *State and community models for improving access to dental care for the underserved*. Chicago: Author.
- American Board of Dental Examiners (ADEX). (April 5, 2005). *Analytical scoring–Endodontics*. Author.
- American Board of Dental Examiners (ADEX). (2006). *2006 Annual Report*. Author.
- American Board of Dental Examiners (ADEX). (2007). *2007 Annual Report*. Author.
- American Board of Dental Examiners (ADEX) (April 10, 2006) *ADEX Examination Committee Report*. Author.
- American Board of Dental Examiners (ADEX). June 23, 2006). *ADEX Examination Committee*.
- American Board of Dental Examiners (ADEX). August 26, 2006). *ADEX Examination Committee*. Author.
- American Board of Dental Examiners ADEX). (April 12, 2007). *Memorandum from Quality Assurance Committee*. Author.
- American Board of Dental Examiners ADEX). (April 17, 2007). *ADEX Examination Committee*. Author.
- American Board of Dental Examiners (ADEX). December 5, 6, 7, 2007). *ADEX Chief Examiner’s Report*. Author.
- American Board of Dental Examiners ADEX). (January 19, 2008a). *Minutes of ADEX Meeting*. Author.
- American Board of Dental Examiners ADEX). January 19, 2008b). *Examination Committee Meeting Minutes*. Author.
- American Board of Dental Examiners ADEX). (January, 22, 2008). *Minutes of ADEX Board of Directors*. Author.
- American Board of Dental Examiners ADEX). (August 21, 2008). *ADEX Quality Assurance Committee*. Author.
- CRDTS (July 12, 2005). Explanation of scoring system. Topeka, KS: Author.
- CRDTS (2008). *Amended and restated bylaws of CRDTS*. Topeka, KS: Author.
- CRDTS (November 8-9, 2008). *CRDTS Dental Examination Review Committee Report*. Topeka, KS: Author.
- CRDTS (August 2009). *CRDTS Dental Examination Committee Meeting*. Topeka, KS: Author.
- CRDTS (April 17-18, 2009). *CRDTS Dental Examination Review Committee Report*. Topeka, KS: Author.
- CRDTS (January 16, 2010). *CRDTS Dental Examination Review Committee Report*. Topeka, KS: Author.
- CRDTS (April 19, 2010). *CRDTS Dental Examination Review Committee Meeting*. Topeka, KS: Author.
- CRDTS (August 26, 2010). *CRDTS Dental Examination Review Committee Meeting*. Topeka, KS: Author.
- CRDTS (October 22, 2010). *Central Regional Dental Testing Services Dental Hygiene Examination Overview*. Topeka, KS: Author.
- CRDTS (October 22, 2010). *2010 New Examiners Orientation*. Topeka, KS: Author.
- CRDTS (2010a). *Chief Examiners Manual*. Topeka, KS: Author

- CRDTS (2010b). *Dental Candidate Manual*. Topeka, KS: Author.
- CRDTS (2010c). *Dental Examiner's Manual*. Topeka, KS: Author.
- CRDTS (2010d). *New examiners orientation*. {Power Point Presentation}. Topeka, KS: Author.
- CRDTS (2010e). *CRDTS' Security Measures*. Topeka, KS: Author.
- Klein, S. P. (April 15, 2008). *Technical Report: Class of 2007–American Dental Licensing Examination*.
- Klein, S. P. (May 9, 2008). How many examiners are needed for case acceptance decisions?
Author.
- Klein, S. P. (November 2, 2010) *Technical Report: Class of 2009–American Dental Licensing Examinations*. Author.
- Ray, L. (March 27, 2008). *Report to CRDTS Examination Review Committee*. Topeka, KS: CRDTS.
- Ray, L. (2010a). *Examiner Profile Statistics–Clinical Restorative Examination*. Topeka, KS.
- Ray, L. (2010b). *Examiner Profile Statistics–Clinical Manikin Examination Endodontics and Fixed Prosthodontics*. Topeka, KS: Author.