CRDTS' National Dental Examination

Technical Report

for the

Year Ending 2017

ACKNOWLEDGMENT

Dr. Thomas Haladyna, Professor Emeritus at Arizona State University, is a well-recognized measurement specialist and educator. An author of many articles and books, Dr. Haladyna has been associated with many projects related to clinical evaluation in the professions. In particular, he was a valuable contributor to the 2004 *AADE Guidance for Clinical Licensure Examinations in Dentistry.* His comprehensive knowledge of the 1999 *Standards for Educational and Psychological Testing* enabled him to cross-reference those *standards* relevant to the principles and guidelines in the AADE Guidance document. In recent years, Dr. Haladyna has begun serving as a consultant to the Equal Employment Opportunity Commission, applying his knowledge of measurement principles to the review of those examinations that may be used as prerequisites for employment.

Dr. Haladyna authored CRDTS' *2010 Technical Report on Clinical Examination in Dentistry,* followed by technical reports in 2014 and 2016 upon the publication of the revised *Standards for Educational and Psychological Testing* in 2015. For all of these technical reports, including this *2017 Technical Report on CRDTS National Dental Examination,* Dr. Haladyna has reviewed all CRDTS' reports, analyses, and examination documents and identified those current *standards* relevant to clinical evaluation and correlated them with the appropriate aspects of CRDTS' *National Dental Examination.*

CRDTS is very pleased that Dr. Haladyna's expertise has documented the validity evidence accumulated in the development, scoring and administration of CRDTS' clinical dental examination. He is a gifted measurement specialist with a knowledge and understanding of clinical evaluation that is outstanding among his professional colleagues.

Lynn M. Ray, RDH, BS
CRDTS Director of Analysis
April 2017

Table of Contents

Introduction

The *National Dental Examination (NDE)* is developed by Central Regional Dental Testing Service (CRDTS) for providing validly interpretable test score information to states and jurisdictions to help each make a licensing decision for those wanting to practice dentistry.

A technical report has the important responsibility of displaying the qualities of a testing program that support validity. This technical report summarizes the argument for validity and the body of evidence supporting that argument. Thus, this technical report contains information useful in evaluating the validity of *CRDTS' NDE* test score interpretation and use.

This technical report is organized in the following way.

1. Validity is defined and discussed as it applies to the *CRDTS' NDE.*
2. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 2015) are applied systematically to support validity.
3. The *CRDTS' NDE* is described.
4. The largest section of this report presents validity evidence in support of using test scores as part of the information used to license dentists.

For improving clarity in this technical report, some terms are defined here.

Standards (2015) refers to the above publication. The *Standards* is a highly respected and well-used set of guidelines for test planning, development, and validation. When *standards* appears in lowercase, this term refers to specific statements in the above publication.

A *test* is a device containing many tasks developed for obtaining test scores. Often the word *examination* or *exam* is used to mean test.

A *testing program* is an organization devoted to designing and developing a test, and validating test score interpretations and any uses. Sometimes the term *examination program* is used as a synonym for testing program.

Validity refers to the reasonableness of interpreting a test score as an indication of a candidate's professional competence. Validity is defined more adequately in a subsequent section of this technical report.

Construct refers to the domain of tasks performed by a dentist. A more recognizable term is *content*. The content of the *CRDTS' NDE* is the construct of professional competence in dentistry. Often we think of content as a domain of tasks, which is called the *target domain*. The target domain is a critical idea in the development of the construct of professional competence in dentistry and the validation of using clinically-based test scores for licensing decisions.

Validity

Validity refers to the judged degree to which an argument and evidence support a specific interpretation or use of a test score. In dentistry, the intended interpretation of the *CRDTS' NDE* test score is how the candidate stands relative to a domain of tasks performed by dentists. The test scores can be used by states to validly determine in conjunction with other information who receives a license to practice dentistry. The test is a representative sample from this domain. This domain is limited to those of normal, everyday practice and not rare or esoteric tasks performed by specialists.

Validation is an investigative process by which the argument and validity evidence can be judged by a competent observer. The judgment is in terms of degrees of validity. Generally, the body of evidence is considered in totality but weaknesses in this body of evidence are noted in this technical report and remedies are advised. This technical report supports an evaluation of this testing program.

For a positive evaluation, the argument has to be sound and compelling, the claims just, and the preponderance of evidence supporting each 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 often recommended (Haladyna, 2006). By studying negative evidence and seeking remedies, validity is increased.

Table 1 shows the constituent elements in validation.

Table 1: Validation of CRDTS's *NDE*	
Argument	The American Dental Association administers the *National Board Dental Examination*. 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' *NDE* is a clinical performance examination intended to measure dental clinical competence directly. These two examinations represent complementary aspects of dental competence. CRDTS's *NDE* is the capstone in this licensing process for licensed dentists.
Claim About Validity	CRDTS claims that candidate scores from its *NDE* 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 technical report provides validity evidence of many types that are based on national test *standards*. CRDTS's documents cited in this report and found in the appendix offer validity evidence supporting this argument.
Evidence Weakening the Argument	CRDTS considers threats to validity and acts accordingly to diminish each threat. By that, CRDTS strengthens the evidence supporting the argument and the claim for validity.
Lack of Evidence	If evidence is missing, CRDTS has the responsibility to gather such evidence in the future as it increases validity.

A Threat to Validity–Construct Representation

The target domain represents a large, ideal set of tasks that licensed dentists typically perform in practice. The size of this target domain is a matter of professional judgment. Administering the entire target domain to a candidate for licensure is impractical. Such a test would entail many days. In some professions, internships are provided so that the candidate for licensure performs many tasks in the target domain under supervision of a faculty member. This is true in dental education. CRDTS claims that its *NDE* represents a fair and sufficient sampling of tasks from the target domain. The domain of tasks was established via a survey of the profession, as is reported subsequently in this technical report. Such a survey is a necessary condition in developing a test like the CRDTS' NDE (Raymond & Neustel, 2006, Raymond, 2016).

Construct representation designates the match between the target domain and actual tasks on the *CRDTS' NDE.* Because a survey of the profession assessed the target domain, CRDTS determined which tasks should be included in its *NDE.* Thus, construct misrepresentation is not perceived as a threat to validity. This technical report provides evidence to support this claim.

Another Threat to Validity–Construct-irrelevant Variance (CIV)

CIV is a technical term for bias. It is systematic error. Such error falsely inflates or deflates a test score. CIV has many sources. For instance, a lenient examiner may overrate a candidate performance. An interruption in test administration may cause a candidate to lose time and fail to perform a task as intended, which results in a deflated score. Testing agencies have a responsibility to identify potential sources of CIV and eliminate or reduce each threat to validity. Throughout this technical report, potential sources of CIV are named, investigated, and reported. As the evidence shows, CIV is NOT a major threat to validity in this testing program.

Integrating Validity Evidence and the Judgment of Adequacy

> "*A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of scores for specific uses*" (*Standards*, 2015, p. 21).

As the *Standards* (2015) state, validation is a never-ending process. This technical report provides a summary of validity at a point of time and offers a historical perspective when compared with previous and subsequent technical reports.

Validity Evidence Used in This Technical Report

To organize validity evidence, the following categories are presented: content, item quality, reliability, examination administration, selection, training, and retention of examiners and scoring, scaling and comparability, standard setting, score reporting, rights of test takers, security, and documentation. This body of evidence is evaluated holistically.

Standards for Educational and Psychological Testing

The *Standards* (2015) update the previous *Standards* (1999). A large, representative committee of testing experts and other qualified volunteers participated in developing these *standards*. For this evaluation, these *standards* are applied and cited in this technical report. All of the referenced *standards* influence the overall judgment of validity. The American Association of Dental Examiners (2003) 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 lists specific *standards* listed here and quoted throughout this report. In each section, a discussion and evidence are offered in support of these *standards*.

Table 2: *standards* Used in this Technical Report	
Chapter 1: Validity. This chapter identifies fundamental concepts and types of validity evidence that appear throughout this evaluation report.	1.0, 1.1, 1.2, 1.5, 1.7, 1.9, 1.11, 1.13
Chapter 2: Reliability. As a primary type of validity evidence, evidence is sought	2.0, 2.5, 2.7, 2.13, 2.14
Chapter 3: Fairness	3.0, 3.1, 3.2, 3.4,
Chapter 4: Test Design & Development	4.0, 4.1, 4.2, 4.3, 4.7, 4.8, 4.10, 4.12, 4.13, 4.16, 4.18, 4.20, 4.21
Chapter 5: Scores, Scales, Norms, Score Linking, and Cut Scores	5.0, 5.1, 5.5, 5.6
Chapter 6: Test Administration, Scoring, Reporting, and Interpretation	6.0, 6.1, 6.4, 6.5, 6.6, 6.8, 6.9, 6.10, 6.14, 6.15, 6.16
Chapter 7: Supporting Documentation	7.0, 7.1, 7.2, 7.4, 7.8, 7.10, 7.13
Chapter 8: The Rights and Responsibilities of Examination Takers	8.0, 8.1, 8.2, 8.6, 8.8, 8.9, 8.10, 8.11, 8.12
Chapter 11: Workplace Testing and Credentialing	11.1, 11.2, 11.3, 11.4, 11.13, 11.14, 11.16

As noted previously, the *Standards* (2015) promotes testing practices that can increase validity. The *Standards* are silent on policy issues. However, policy decisions can be informed by technical reports that consider the *Standards*.

The *Standards* have some important disclaimers:

1. Not all *standards* apply to a specific testing program. Thus, evidence need not be presented for every standard. In this technical report, *standards* were selected that bear on validity for a clinical performance test used as part of the criteria for licensing dentists in states and other jurisdictions.

2. If there is a legal challenge to a test score interpretation or use, *standards* provides a valuable and reputable basis for understanding and defending against a challenge. CRDTS

can use *standards* as a basis for its credibility if legal challenges were made on a test score interpretation or use.

Throughout this technical report, *standards* are quoted that apply to this testing program and relate to validity. Thus, readers are encouraged to consider that (1) *standards* are followed in test design, development, administration and scoring, and (2) the application of these *standards* with proper documentation in this report increases validity.

Description of the *CRDTS' National Dental Examination (NDE)*

The best current source of information about this testing program comes from CRDTS' website: CRDTS.org. Detailed information about the examination can also be found in the 2017 *Dental Candidate Manual* (CRDTS, 2017b). The appendix of this technical report contains many archived documents that attest to the development of the *CRDTS' NDE* and validation of interpretation and use of test scores. A recent Dental Examination Review Committee meeting (CRDTS, January 2017a, 2017b, June 3, 2017) describes some changes in the examination and examination procedures. These changes included clarifications, updates, and the implementation of a dress code for examiners.

To be licensed to practice dentistry in any state or United States' jurisdiction, a candidate has to meet many qualifications, one of which include passing a series of tests. The *National Dental Board Examination–NDBE* (Parts I and II) is one 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 known as the *CRDTS' NDE*. CRDTS was established in 1972. As stated in its bylaws, state boards for dental licensing are its members. Its members meet annually in August.

The *CRDTS' NDE* is used to measure a candidate's clinical competence in dentistry in four distinct areas of dentistry. 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. The cut score for each test is 75 for making pass/fail decisions. The cut score is set by legislation in participating states.

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.

Origin of Current Examination

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; June 23, 2006; August 26, 2006; April 12, 2007; April 17, 2007; December 5-6-7, 2007; January 19, 2008a; January 19, 2008b; January 22, 2008; August 21, 2008).

As of June 30, 2009, CRDTS severed its association with ADEX but retained much of the examination design and structure. CRDTS had actively participated during its development over a four-year period. One report by ADEX (January 10, 2008) provides an example of examination review and recommendations that bear on the current examination. Up to that point, documentation of validity was done by ADEX. After that time, the responsibility for subsequent documentation and any modifications of the examination has been the responsibility of CRDTS.

Traditional versus Curriculum Integrated Formats

All qualified candidates for licensure currently enrolled in dental schools have the option of taking the *CRDTS' NDE* in the Curriculum Integrated Format, which allows earlier administration of examinations with the caveat that if a candidate fails, remediation and retesting are available for them while they are still under faculty supervision. The traditional format requires that candidates take all examinations at the end of their dental education. No material difference exists in the content or difficulty of the *CRDTS' NDE* in either format.

Conjunctive Versus Compensatory Scoring

For any test, the test agency can require candidates for licensure to pass each test in a series of tests. This requirement is known as *conjunctive scoring*. Conjunctive scoring is a high standard, because poor performance in any test is not tolerated and results in failure. The purpose of a licensing examination is to screen candidates who may practice unsafely and harm patients. Thus, 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 many (see Haladyna and Hess, 1999). Responsible boards prefer conjunctive scoring because of the safeguards it provides.

Compensatory scoring requires that a pass/fail decision be made on the total scores from all tests in a series. Low performance in one area can be made up by higher performance in another area. All the candidates have to do is to earn a total score high enough to meet or exceed a single cut score.

Compensatory scoring is more lenient than conjunctive scoring. 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 uses conjunctive scoring for all four examinations. However, compensatory scoring is applied with those tests having multiple procedures within their content. A candidate may perform lower in one procedure, but if their performance is above average in the other related skill sets in that test section, they may achieve a passing score. The rationale for conjunctive scoring follows a line of reasoning that asserts that low performance in any of the four content areas is unsatisfactory. 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 via legislative action.

***CRDTS' NDE* Structure and Content**

As noted previously, the *CRDTS' NDE* consists of four tests. Each test is scored on a 100-point scale: Endodontics, Prosthodontics, Periodontics, and Restorative. The *CRDTS' NDE* has been developed and refined over many years in consultation with subject-matter testing experts

(SMEs). A practice analysis survey of the profession is an important periodic step in verifying the content of the test (Raymond, 2016). Each year, improvements are made in the examination that improve validity. Past technical reports and other cited documents trace the history and continued improvement of this testing program.

A generic scoring guide is used by three highly trained, skillful dental examiners to rate the performance on a variety of tasks. The rating scale has four identifiable performance levels: (1) Satisfactory, (2), Minimally acceptable, (3) Marginally substandard, (4) Critically deficient.

Endodontics

It is a manikin-based test. It consists of two procedures: an access opening on an artificial posterior tooth and an access opening, canal instrumentation and obturation on an artificial anterior tooth. The criteria for these procedures are combined and scored in total using 17 criteria: anterior–12 and posterior–5. Total score is 100 points. Penalty points may be assessed. The computation for total score is the ratio of points earned and points possible multiplied by 100.

Fixed Prosthodontics

It is a manikin-based test that includes three procedures:

1. Preparation of tooth #5, a single-layered artificial tooth, for a porcelain fused to metal crown as one abutment for a 3-unit bridge.

2. Preparation of tooth #3, a single-layered artificial tooth, for a cast gold metal crown as the other abutment for the same 3-unit bridge.

3. Preparation of tooth #9, a single-layered artificial tooth for a full ceramic crown.

The titles and number of criteria for scoring are:

	Subtest (procedure)	Criteria (Items)
1	Cast gold crown	10
2	Porcelain-fused-to-metal crown preparation	10
3	Ceramic Crown Preparation	11

Penalty points can be assessed.

Periodontics

The components of the periodontics examination are listed below. Points may be deducted for treatment selection and/or treatment *standards*.

Component Name and Abbreviation	Items	Points/ Item	Total Points
Extra/Intraoral Assessment	8	2.00	16
Periodontal Measurements/Gingival Recession	16	0.75	12
Scaling/Subgingival Calculus Removal	12	5.00	60
Supragingival Deposit Removal	12	2.00	12
Total			100

Tissue Management involves penalty points. Five points are deducted for each confirmed error. If three confirmed errors occur, a critical failure results.

Restorative

This test is patient-based. Candidates are required to complete a Class II and a Class III procedure. For Class II, candidates must take test 1–Anterior Composite Preparation and Finish. After taking test 1, they can choose from the remaining three. Most choose Posterior Composite and Finish.

Four Tests–Two Subtests for Each Test	Items
1a. Anterior Composite Preparation	8
1b. Anterior Composite Finish	9
AND	
2a. Posterior Composite Preparation	11
2b. Posterior Composite Finish	8
OR	
3a. Amalgam Preparation	13
3b. Amalgam Finish	7
OR	
4a. Class II Slot Preparation	9
4b. Class II Slot Finish	8

The total score is the percentage of points earned divided by points possible multiplied by 100. The scale ranges between zero and 100. Points can be deducted for critical deficiencies as determined by the examiners.

VALIDITY EVIDENCE BEARING ON ALL FOUR TESTS

Because the *CRDTS' 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. Then, four sections follow this section. Each of these sections provides evidence that is unique to that test.

Validity

The *standards* cited here deal directly with validity. Content is a major type of validity evidence. Table 3 lists the *standards* that directly apply to validity. Some *standards* are quite lengthy, so they were paraphrased and presented in italics.

Table 3: *Standards* Generally Related to Validity	
1.0	Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided.
1.1	The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.
1.2	A rationale should be presented for each intended interpretation of test scores for a given use together with a summary of the evidence and theory bearing on the intended interpretation.
1.5	When it is clearly stated or implied that a recommended test score interpretation for a given use will result in a specific outcome, the basis for expecting that outcome should be presented together with relevant evidence.
1.7	If test performance, or a decision made therefrom, is claimed to be essentially unaffected by practice and coaching, then the propensity for test performance to change with these forms of instruction should be documented.
3.0	*Construct-irrelevant variance (CIV) should be avoided in all aspects of test development, administration, scoring, and reporting.*
3.1	Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.
3.2	Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical or other characteristics.
3.4	Test takers should receive comparable treatment during the test administration and scoring process.
4.0	Tests and testing programs should be designed and developed in a way that supports validity of interpretations of test scores for their intended uses.
4.13	When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.

6.0	To support useful interpretation of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected.
11.1	*A clear statement of intended interpretation of a test score and the use to which it is intended should be made clear to test takers.*

Some *standards* may seem repetitious because each chapter was developed by different sets of testing experts. Thus, this repetition emphasizes the importance of several qualities found in this testing program and evident in this report:

1. Competence in dentistry is defined by a target domain of tasks.

2. A practice analysis is conducted regularly to ensure that the content of each of the four tests has high fidelity with this domain of tasks.

3. All aspects of test development are refined and well described in this technical report and other documents referenced in the appendix.

4. Threats to validity are regularly investigated, and attempts are made to reduce or eliminate these threats.

1. Content

The most fundamental type of validity evidence for a credentialing examination is content-related (Kane, 2006). A dental clinical examination should identify a domain of tasks performed by a competent dentist. Ideally, the tasks in the target 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, 2016; Raymond & Neustel, 2006). As quoted directly:

> *An investigation of a certain occupation or profession to obtain descriptive information about the activities and responsibilities of the occupation or profession and about the knowledge, skills, and abilities needed to engage successfully in the occupation or profession* (*Standards*, 2015, p. 222).

Table 4 presents *standards* bearing on content.

Table 4: *Standards* Related to Content-related Validity Evidence	
1.11	*The basis for defining and identifying content should be clearly specified.*
1.13	If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided.
1.14	When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided.
4.1	Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).
4.2	*Test specifications should be very comprehensive regarding content, test length, item formats, ordering of items and sections, and administration time.*
4.3	*All test development activities should be documented.*
4.12	Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.
5.1	Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations.
11.2	Evidence of validity based on test content requires a thorough and explicit definition of the content domain of interest.
11.3	*When test content is a primary source of validity evidence, a close link between test content and the profession being assessed is required.*
11.13	*The content domain should be clearly described and justified in light of the professional competency being tested.*

Chapter 11 of the *Standards* (2015) is devoted exclusively to *standards* affecting licensure examinations, such as CRDTS's. Not only is 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 the practice analysis. *Standards* 11.2, 11.3, 11.13 all address slightly different but complementary aspects of practice analysis as a basis for test specifications. The test specifications guide examination development.

Practice Analysis (Also known as Occupational Analysis or Job 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, 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.

In 2013, an occupational analysis was done to update the content of the *CRDTS' NDE*. It was concluded: "*Based on these findings, there appears to be no basis for substantive changes to the examination content of clinical procedures*" (CRDTS, 2013, p. 12). Thus, no changes were made in the content of the *CRDTS' NDE*. However, note that regularly scheduled Dental Examination Committee Meetings will make minor changes in items that in no way reflect content changes but may slightly alter the meaning of some items. The criterion by which we judge such alterations is subject-matter expertise. If committee members unanimously agree in an item alteration that does not affect the content domain, then the action is defensible as to content-related validity evidence. This action is not a threat to validity.

Structural Evidence

Is dental clinical competence a single entity consisting of highly related tasks? Or is competence a family of independent tasks, each of which is important in achieving a satisfactory level of competence? Table 5 provides descriptive statistics for Endodontics, Prosthodontics, Periodontics, and Restorative. These test score characteristics for each test affect the estimation of reliability and the estimation of the degree of random error found in test scores. As the examinees are highly trained and highly skilled, we would expect performances on each of the four tests to be very high. Thus, these results in no way suggest that the test be too easy or the scores have a large degree of random error.

Correlations among these four tests range from 0.05 to 0.14. Although, these correlations are statistically significant ($p < 0.05$), the magnitudes are very small. Cronbach's alpha is a measure of internal consistency. If the four tests were combined into a single test, alpha would be 0.27. This is very low, and this result is further indication of the independence of these four tests. A correct interpretation of these results is that the four tests are very independent.

Table 5: Descriptive Statistics for the Four Tests of the CRDTS' NDE				
	Endodontics (II)	Prosthodontics (III)	Periodontics (IV)	Restorative (V)
Candidates	723	684	692	677
Low Score	0	75	75	75
High Score	100	100	100	100
Mean	91.58	92.79	96.11	95.66
Stand. Dev.	20.24	4.60	5.57	3.73
Skewness	-4.09	-1.11	-1.83	-1.58

Claim Supporting Validity

A practice analysis confirmed the content of the *CRDTS' NDE* (CRDTS, 2013). The study of structure also shows the four tests are independent. CRDTS has determined that the four tests are independent enough and important to clinical practice to demand that candidates pass each of these four tests of dental competence sub-abilities. This evidence supports a conjunctive strategy (four independent tests) as opposed to a compensatory strategy (combining four tests into one).

2. Item Quality

The kinds of test item formats used in any testing program can vary significantly (Haladyna and Rodriguez, 2014). These formats include performance, multiple-choice, objective-structured clinical examination, laboratory exercises, manikin tasks, chart-stimulated evaluation, longitudinal, repeated observations, and portfolio to mention a few. No matter the specific formats employed, a rationale must be provided that shows that each test item elicits the desired behavior for a specific task in the domain of relevant tasks defining the profession. Each task on each of the four tests should connect directly to the practice analysis results. Professional judgment by highly qualified, licensed, experienced dentists is crucial to supporting item development and validity.

The *Standards* (2015) are very explicit about the role of item development in test development and validation. Table 6 lists relevant *standards* for item development.

Table 6: *Standards* Related to Item Quality	
4.7	The procedures used to develop, review, and try out items and to select items from the item pool should be documented.
4.8	The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.
4.10	*Statistical properties of item scores should be studied in an appropriate theoretical context.*

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 & Rodriguez, 2014), because the item undergoes the same procedure of validation as we do for test scores. Thus, the evidence needed to conclude that the items used in this examination have been validated include the following:

1. Practice analysis identifies the knowledge, skills, and abilities needed to practice safely and competently.
2. Test specifications are created that identify this content.
3. Items are developed to match the test specifications.
4. Items undergo intensive review by SMEs on content subcommittees.
5. The scoring procedure is developed and is assigned a 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.

Evidence concerning item development comes primarily from regularly scheduled CRDTS Dental Examination Review Committee Reports. As listed in the appendix, these reports provide abundant detail of item development and validation (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, January 15-16, 2011; April 29-30, 2011; August 25, 2011; January 14-15, 2012; April 27-28, 2012; August 23, 2012; January 12-13, 2013; April 26-27, 2013; August 22, 2013; January 11-12, 2014; April 12, 2014; August 22, 2014; January 17, 2015; April 11, 2015; January 9, 2016; April 30, 2016; January, 2017a; January 2017b; June 3, 2017). These committee meeting reports are highly detailed and show the attention given to the continuous improvement of test items. Noting that the test items include tasks to be performed is important and highly complex scoring protocols that require extensive examiner training and reliable judgment. The results of this training and judgment are reported elsewhere in this technical report.

Fidelity

As noted previously, tasks on any clinical performance 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 *CRDTS' 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 identical or similar to tasks performed by dentists on actual patients in dental practice, or, with manikin-based testing, the tasks performed must have high fidelity with actual patient practice. The previously cited committee reports and the practice analysis provide evidence of fidelity (CRDTS, 2013).

Weighting of Test Items

This topic is very important as weights assigned to items have consequences for candidates' scores. In the development of the *CRDTS' NDE*, ADEX and CRDTS have 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. The weighting of any test item is a matter of professional judgment by SMEs. 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. A useful reference is the *Dental Examination Candidate Manual* (2017b), which is available publicly on its website (CRDTS.org).

Claim Supporting Validity

The claim is made in this technical report that item development meets high *standards* as described in various sources including the *Standards* (2015), the first edition of the *Handbook of Test Development* (Downing & Haladyna, 2006), the second edition of the *Handbook of Test Development* (Lane, Raymond, Haladyna, 2016), and *Developing and validating test items* (Haladyna & Rodriguez, 2014).

3. Reliability

Every test score has an unknown degree of random error. This error can be positive or negative and large or small. There is no way to discover how much random error is in a test score. For a candidate whose test score is at or near the cut score of 75, 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). To rephrase this state of affairs, one candidate who exceeds 75 may have a negative random error resulting in a failing decision. Another candidate who scores below 75 due to random error falsely passes. We call these classification errors Type I and Type II. Reliability affords us understanding of the risk of misclassifying candidates whose true scores are at or 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.

Several *standards* apply to reliability and are presented in Table 7:

Table 7: *Standards* Related to Reliability	
2.0	Appropriate evidence of reliability/precision should be provided for the interpretation and use for each intended score use.
2.2	The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretation for the use of test scores.
2.5	Reliability estimation procedures should be consistent with the structure of the test.
2.7	Inter-judge and intra-judge consistency of ratings should be studied, monitored, and documented.
2.13	The standard errors of measurement, both overall and conditional (if reported), should be provided in units of each reported score.
2.19	*Method of opinion of reliability should be documented.*
11.14	Estimates of the consistency of test-based credentialing decision should be provided besides other sources of reliability evidence.

CRDTS have taken steps to ensure to maximize reliability and minimize the risk of misclassification.

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. Also reports by Ray and Cobb (2017b, 2017c, 2017d) report characteristics of the ratings.
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, 2017c).

Conventional reliability estimation depends on high internal consistency among item responses. That is to say, item responses tend to be highly intercorrelated. Sometimes, a clinical performance test can consist of tasks that are not highly related. In this instance, 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 negatively skewed (see Table 5). Statistical techniques, such as reliability and correlation depend on a normal distribution of test scores with considerable variation in test scores. CRDTS' test scores are very restricted due to high performance. Thus, reliability estimates tend to be very low because of skewness in scores. However, random error is also low. So the problem of reliability estimation is ameliorated because 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. Consequently, states using test scores as part of a licensing decision 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 status of candidates whose observed scores fall at or near the cut score of 75. 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.

Reliability results are reported in appropriate sections of this report. In these sections, stratified alpha is used. Sometimes, stratified alpha is high enough that when considering the variation in test scores, the margin of error is small. The number of candidates observed close to the cut score are few or none.

Claim Supporting Validity

The frequency of observations by examiners and the use of well-trained examiners to achieve consistency in ratings makes for highly reliable test scores. This result in turn makes the standard error of measurement around the cut score minimal. Thus, few candidates are in jeopardy of being misclassified. Data is reported in sections devoted specifically to each test in this technical report that shows the small degree of risk of misclassification due to random error.

4. Examination Administration

Test administration is an important aspect of any testing program. McCallin (2006, 2016) provides a very detailed account of issues in examination administration and potential threats to validity. The *Standards* (2015) also provides guidance as several *standards*, shown in the table below.

Table 8: *Standards* Related to Test Administration	
4.16	The instruction presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each format or major area in the test's classification or domain should be provided to the test taker prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.
6.1	Test administration should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instruction from the test user.
6.4	The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.
6.5	Test takers should be provided appropriate instructions, practice, and other support necessary to reduce construct-irrelevant variance.
6.6	Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means.
6.7	Test users have the responsibility of protecting the security of test material at all times.

This standardized examination has been administered over many years. The examination administration has been improved annually. When ADEX was responsible for the examination development, 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, its Examination Review Committee was reactivated and its subcommittees met regularly to improve examination administration (CRDTS, November 8-9, 2008; August 2009; April 17-18, 2009; January 16, 2010; April 19, 2010; August 26, 2010; October 22, 2010; January 15-16, 2011, April 29-30, 2011; August 25, 2011; January 14-15, 2012; April 27-28, 2012; August 23, 2012; January 12-13, 2013; April 26-27, 2013; August 22, 2013; January 11-12, 2014; April 12, 2014; August 22, 2014; January 17, 2015; April 11, 2015; January 9, 2016; April 30, 2016; January 2017a; January 2017b; June 3, 2017).

Another useful source of information about administration is the *Dental Examiner's Manual* (2017c). This 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 manikin and patient-based procedures and each of the four tests in this examination program.

Another useful document is the *Chief Examiner's Manual* (2017a). This notebook contains more than 100 pages of information about the role of the chief examiner from three months before

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.

Claim Supporting Validity

The examination administration is very well organized and standardized. This testing program has reached a high level of proficiency in examination administration as evidenced in the cited documents and as described by McCallin (2006, 2016).

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

Table 9 lists *standards* related to selection, training, and retention of examiners. Also, *standards* in this table addresses scoring. The development of the scoring system is documented in a report (CRDTS, July 12, 2005).

Table 9: *Standards* Related to Scoring	
1.9	*When candidate performance is judged, the process for identifying, recruiting, training, and monitoring judges should be documented.*
2.7	*Inter-judge and intra-judge consistency of ratings should be studied, monitored, and documented.*
4.18	Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed-responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.
4.20	*Processes for identifying, training, and evaluating judges should be well developed and documented.*
4.21	*Rater consistency and rater effects should be studied, documented, and, if feasible, improved.*
5.0	Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed uses.
6.8	Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring.
6.9	Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic errors should be documented and corrected.

Of great importance to validity is the way examiners are identified and chosen, how they are trained, how they are evaluated, and then retained or dismissed. Additionally, we have several issues involving scoring. According to the *Dental Examiner's Manual* (CRDTS, 2017c), CRDTS examiners are required to calibrate before every examination, no matter how recently they may have examined.

For any given examination, examiners are given specific assignments as Chief, Team Captains, Clinic Floor Examiners or scoring examiners. The Chief Examiner leads the examiners through a general calibration and a Team Captain presents patient acceptability calibration. Clinic Floor Examiners then break away for a separate calibration while the scoring examiners participate in a four to five-hour calibration led by the Restorative and Periodontal Team Captains, respectively. The calibration exercises for each team are custom-designed PowerPoint presentations with an Audience Response system so the examiners can independently evaluate case situations, click their responses, and see variation on the screen when all responses are submitted. A script and post-test are provided for the Team Captains as well. The calibration exercises are designed and updated annually by subcommittees of the Examination Review Committee.

For many years, CRDTS assigned separate specialty teams for Restorative and Periodontics, due to the substantial variation in the evaluation protocol for both exam sections. With the progressive refinement of CRDTS' electronic scoring system to streamline both administrative and exam protocols, CRDTS is now cross-training examiners to evaluate all patient-based procedures. The criteria for every scorable item is presented systematically on the screen of the tablet as the examiners progress through each evaluation. For the Periodontal evaluation, the tablet combines the examiners' findings to decide if the treatment selection meets the criteria, and verifies that the surfaces the examiners select for the final evaluation fulfill the established protocol. Any issues that arise are resolved by the respective Team Captains. This greatly reduces any incidence of inconsistency and makes more efficient use of examiners' time and expertise. In addition, efficiency is enhanced by the Flow Management System used throughout 2015 TO 2017. From the moment a patient is checked in for evaluation, the computer tracks the timing, and as soon as an examiner has finished one evaluation, the tablet tells them which operatory they should go to next, and when they get there, their information is already loaded in that tablet. They are ready to start the evaluation. Examiners can refuse an assignment for personal privilege, and the assignment is immediately transferred to another examiner. When an examiner returns to the evaluation center, they can put themselves back in the assignment pool. The system also tracks each examiner's frequency of assignment and their evaluation time. Examiners have responded well to the efficiency of the system, and CRDTS is provided a log of the entire examination.

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, 2017c). This system includes home study and onsite opportunities for examiners. An 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. Documents describing the performances of examiners are an important aspect of diagnostic feedback (Ray & Cobb, 2017b, 2017c, 2017d).

> *The CRDTS Examination Review Committee maintains an "Examiner Profile Service" specifically designed to provide each CRDTS examiner with information which the Committee hopes individual examiners will use to self-assess and, where necessary, improve their individual examining skills.* (Ray & Cobb, 2017a).

Examiner calibration exercises are provided online as well as prior to each examination. All examiners participate in team meetings and work as teams. When scores have been released from a particular testing site, examiners receive an online graph of how their average scores on each procedure compare with their teammates who evaluated the same cases. At the end of each examination year, all examiners are given a profile of their performance criterion by criterion. This information is used for remediation of examiners and for retention.

Scoring

Extensive committee work was reported on all aspects of scoring (CRDTS, November 8-9, 2008; August 2009; April 17-18, 2009; January 16, 2010; April 29, 2010; August 26, 2010; October 22, 2010; January 15-16, 2011, April 29-30, 2011, August 25, 2011; April 27-28, 2012, August 23, 2012; January 12-13, 2013; April 26-27, 2013; August 22, 2013; January 11-12, 2014; April 12, 2014; August 22, 2014; January 17, 2015; April 11, 2015; January 9, 2016; April 30, 2016; January, 2017a; January, 2017b; June 3, 2017). The *Dental Examination Candidate Manual* (CRDTS, 2017b) 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 onsite and ratings are recorded electronically. 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 examiners are subjected to a multi-step process for standardization and calibration 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 Examination Candidate Manual* is also revised (CRDTS, 2017b).

After the examinations are administered, CRDTS annually produces reports of examiner performance, which are intended for examiner self-assessment (Ray & Cobb, 2017b, 2017c, and 2017d). These results are also used to evaluate examiners and to inform decision-making for future examiner assignments. Such information can be very useful in refining training and improving examiner consistency or, if justified, removing examiners who are inconsistent. 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, decides their effectiveness, looks for emerging leadership qualities as Team Captains or Chief Examiners, and assigns them as scoring examiners or Clinic Floor Examiners 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 examination. 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 end 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.

Claim Supporting Validity

The training of examiners by CRDTS is a highly refined activity that has received considerable attention over many years. The *Dental Examiner's Manual* (CRDTS, 2017c) is an annual publication updated each year. It contains comprehensive information related to training and scoring. This document is supplemented with other materials used during training. Thus, training of examiners and their scoring is very effective.

6. Scaling & Comparability

Chapter 5 of the *Standards* (2015) is devoted to scaling and comparability. Table 10 list *standards* related to this important topic.

Table 10: *Standards* Related to Scaling and Comparability	
5.2	The procedures for constructing scales used for reporting scores and the rationale for these procedures should be clearly described in detail.
5.5	When raw scores or scale scores are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretations should be explained clearly.
5.6	Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of scale on which scores are reported.

The validity of interpreting test scores is strongly dependent on having a test score scale that is constant from one examination administration to another. Considering that the cut score is also constant, it is important that the test be equally difficult and the content fixed on all occasions it is administered. With multiple-choice tests, it is customary to have multiple forms that have to be equated so the scale is constant from one test form to another test form. With a clinical performance test, scaling and comparability of results are very different. There is only one test form. It is administered many times in many different places. The tasks are well known, and all candidates have an equal opportunity to prepare for the test. Because the tasks are those that licensed dentists must perform competently, there is complete transparency between the target domain and the test representing this target domain.

Thus, with the *CRDTS' NDE*, the only variable is the set of examiners for any one test. All examiners come from a common pool of examiners. All are highly qualified and extensively trained. Their ratings are calibrated before they rate performance. 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 reported in subsequent sections of this technical report. Table 11 reports failure rates for the past three years. Some stability in these rates and some variability exists. Many factors may contribute to variation including training in dental schools, demographic differences, and number of candidates. Reports of rater consistency and bias in ratings does not seem to be a threat to validity here (Ray & Cobb, 2017b; 2017c; 2017d).

Table 11: Failure Rates for Three Successive Years					
	Restorative	Endo.	Pros.	Perio.	Candidates
2014	11.6%	6.7%	10.2%	3.4%	674
2015	9.5%	6.8%	9.2%	3.7%	816
2016	9.6%	6.5%	8.1%	3.2%	786
2017	9.5%	8.2%	9.9%	2.8%	905

Claim Supporting Validity

Scaling for comparability appears adequate given that this is a clinical performance test where the tasks are well known and frequently practiced by candidates. The use of three examiners helps stabilize the scale, which are also 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 from one administration to another.

7. Standard Setting

For the four CRDTS tests used to make pass-fail decisions, a cut score is established. Cut scores are set by states. Generally, these states have adopted a cut score of 75.

Table 12 below lists four relevant *standards*. This section provides evidence relating to these four *standards*. Note that these *standards* tend to be repetitious because they come from different sources and are found in different chapters in the *Standards* (2015).

Table 12: *Standards* Related to Setting the Cut Score	
5.5	When raw scores or scale scores are designed for criterion-reference interpretation, including the classification of examinees into separate categories, the rational for recommended score interpretations should be explained clearly.
5.21	When proposed test score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.
5.23	When feasible and appropriate, cut scores defining categories with distinct substantive interpretation should be informed by sound empirical data concerning the relations of test performance to the relevant criteria.
11.16	The level of performance required for passing a credentialing test should depend on knowledge and skills necessary for credential worthy performance in the occupation or profession and should not be adjusted to control the number or proportion of persons passing the test.

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. Most candidates taking this test far exceed the cut score. A panel of SMEs have reviewed a state's mandate and performance of candidates relative to this cut score. They have recommended that the cut score of 75 is appropriate for making a pass/fail decision.

Claim Supporting Validity

Because states mandate cut scores, CRDTS' SMEs have reviewed and endorsed the cut score. Thus, the way cut scores were established meet these *standards*.

8. Score Reporting

The *Standards* (2015) also addresses issues related to score reporting. Table 13 below shows *standards* addressing this topic.

Table 13: *Standards* Related to Score Reporting	
6.10	When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what the scores represent, the precision/reliability of the scores, and how scores are intended to be used.
6.14	*Test organizations should maintain confidentiality and protect the rights of test takers.*
6.15	When individual test data are retained, both the test protocol and any written report should also be preserved in some form.
6.16	Transmission of individually identified scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores and pertinent ancillary information.

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 Periodontics 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. Candidate score reports are treated confidentially.

Additionally, CRDTS provides an annual report on the performance of graduates from each dental school taking the *NDE* (Ray & Cobb, 2017a). This report provides a basis for formative evaluation by which each school can identify strengths and weaknesses in the performances of its graduates and, by that, improve their instructional and training program.

Claim Supporting Validity

CRDTS meets these standards.

9. Rights of Test Takers

Chapter 8 of the *Standards* (2015) is devoted to the rights of test takers. Table 14 below lists *standards* relevant to the rights of test takers.

Table 14: *Standards* Related to the Rights of Test Takers	
8.1	Information about test content and purposes that is available to any test taker prior to testing should be made available to all test takers.
8.2	Test takers should be provided in advance with as much information about the test, the testing process, the intended use, test scoring criteria, testing policy, availability of accommodations, and confidentiality protection as is consistent with obtaining valid responses and making appropriate interpretation of test scores.
8.5	*Policies for release of test scores should be carefully considered and clearly recommended. Release of scores should be consistent with the purpose of the test and in consideration of the test takers and informed consent.*
8.6	*Transmission of test taker scores should be protected from improper use.*
8.8	*When test scores are used to make decisions, the test taker should have access to that information.*
8.9	*Test takers should be aware of the consequence of cheating.*
8.10	*In the instance of an irregularity, a test taker should be informed of any delay in score reporting.*
8.11	*In the instance where a test result is invalidated, the test taker must have access to all information bearing on that decision. Ample opportunity should be available for appeal and claims.*
8.12	*Test takers are entitled for fair treatment in the event of a irregularity that prevents a score from being reported or if a score is invalidated. Test takers should have a means for recourse of any dispute regarding the rejection of a test score for a decision.*

The *Dental Examination Candidate Manual* (CRDTS, 2017b) contains many topics important to candidates. CRDTS website supplies information about application and eligibility, the calendar for administration, examination content, scoring, forms and manuals, online application, and orientation. 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 re-score the examination, but will consider any appropriate evidence that points to alternative results. 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 Examination Candidate Manual*.

Claim Supporting Validity

These *standards* are met. The *Dental Candidate Manual* (2017b) is the best source of information supporting this claim.

10. Security

CRDTS has taken many steps to ensure security in examination development, administration, scoring, and reporting (CRDTS, 2014c). The following *standards* apply to security.

Table 15: *Standards* Related to Security	
6.7	*Test users have the responsibility of protecting the security of test materials at all times.*
6.14	*Testing organizations should have a safe, secure system to store test information.*
6.15	When individual test data are retained, both the test protocol and any written report should also be preserved in some form.
6.16	Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores and pertinent ancillary information.

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, 2014c).

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.

Claim Supporting Validity

CRDTS has a well-developed system ensuring security in all phases of examination planning, development, administration, scoring, and reporting.

11. Documentation

Chapter 7 of the *Standard*s (2015) states:

> *"The objective of the documentation is to provide test users with the information needed to help them assess the nature and quality of the test, the resulting scores, and the interpretations based on the test scores"* (p. 123).

The table below provides *standards* related to documentation. Most of the *standards* in this table duplicate other *standards* throughout this report. The important consideration here is that CRDTS has ample documentation that will be presented in this section that also fulfills these *standards*.

Table 16: *Standards* Related to Documentation	
7.0	Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores.
7.1	The rationale for a test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. When particular misuse of a test can be reasonably anticipated, cautions against such misuses should be specified.
7.3	When the information is available and appropriately shared, test documents should cite a representative set of studies pertaining to general and specific uses of a test.
7.4	Test documentation should summarize test development procedures, including descriptions and the results of the statistical analyses that were used in the development of the test, evidence of the reliability/precision of scores and the validity of their recommended interpretations, and the methods for establishing performance cut scores.
7.8	Test documentation should include detailed instructions on how a test is to be administered and scored.
7.10	Tests that are designed to be scored and interpreted by test takers should be accompanied by scoring instructions and interpretive materials that are written in a language the test takers can understand and that assist them in understanding the test scores.
7.13	*Supporting documents should be made available to the appropriate people in a timely manner.*

As noted in the appendix, CRDTS has a large collection of documents attesting to meetings, publications, manuals, studies, and reports bearing on test development and validation.

Claim Supporting Validity

Throughout this technical report, these documents are cited in reference to *standards*. By that, it is argued that validity is served and improved. The annual technical report alone stands as a single authoritative source of validity evidence matched to *standards*.

VALIDITY EVIDENCE BEARING ON SPECIFIC TESTS

This final section of the technical report focuses on three important pieces of validity evidence: structure of data, examiner consistency and reliability for each of the four tests.

For examiner consistency, the following will show that three examiners for each examinee task have a high degree of consistency. The percentage reports in subsequent tables are for perfect agreement. Often examiners will have a one point difference. For instance: examiner A-3, examiner B-3, examiner C-4. Although there are one perfect agreement and two imperfect agreements, the candidate receives a rating of three, which is the median. Thus, imperfect agreement does no harm to a candidate score because the median is used instead of the mean.

For reliability, the matter is more complex. First, because performance is high, few candidates have scores near 75 (the cut score for pass/fail decisions). Thus, the risk of being misclassified as a pass or fail is very small. Second, because of the negatively skewed distribution, reliability estimates tend to be low. However, the margin of error (standard error of measurement) is also small due to the small variance of test scores. Finally, some subscales of each test appear to be independent of other subscales. Thus internal consistency type reliability estimates tend to be lower than they should be. In these instances, stratified alpha was used (Haertel, 2006; Nunnally & Bernstein, 1994).

Thus, reliability estimates involve stratified alpha and refer to complete data sets where candidates have a complete set of subscores. Scores of zero on a subtest lead to automatic failure when using the compensatory scoring model because the mean of the subscores is typically less than 75.

Endodontics Test

This test comprises two parts: anterior and posterior endodontics. Penalty points are part of scoring.

Structure of the Data

Correlation between the anterior and posterior scores was a very low 0.17. Although this result is statistically significant, the degree of relationship is practically zero. This result is also influenced by the fact that performance on these two measures by all candidates was consistently high. This is not surprising as candidates for a licensing examination who have been well-trained should perform consistently high. CRDTS argues that the anterior and posterior measures are vital to the measurement of endodontic proficiency.

Examiner Consistency

Ray and Cobb (2017b) report examiner consistency in a variety of ways. Confirmed scores accuracy is 91.27. The method of scoring that uses the median reduces random error. Considering the more consequential pass/fail analysis, examiner consistency is extremely high.

Reliability

Table 17 reports descriptive statistics and reliability estimates for each subscale and for the total scorer using stratified alpha.

Table 17: Descriptive Statistics and the Reliability Estimate

	N	Items	Points	Mean	S. D.	Skew	Rel.
Anterior–12 sets of observation	758	12	48	46.8	2.1	-2.4	0.46
Posterior–5 sets of observations	758	5	20	17.8	2.9	-1.4	0.55
Raw Score	758	17	68	64.6	3.9	-1.4	0.58
Scaled to 100 points.	758	17	100	95.0	5.7	-1.4	0.58

Standard Error of Measurement

The margin of error surrounding the cut score of 75% is 3.7 on the 100-point scale. The margin of error (one standard error of measurement) can be constructed around the cut score. That zone ranges from 71.3 to 78.7. So those candidates whose score are 72 to 77 are caught in this zone of uncertainty and are in risk of misclassification due to random error. Of the 758 candidates included in this analysis, only 19 fall in this zone of uncertainty around the cut score. As these candidates have performed poorly in comparison with most candidates, the risk of misclassification is associated with their poor performance. There is very little any testing

program can do to reduce this zone of uncertainty. The three candidates failing as a result are counseled to seek remediation and re-take the examination.

Claim Supporting Validity

Examiner consistency is very high for the anterior endodontics and lower for the posterior endodontics. The reliability estimate for the total score is moderate, but this fact is mitigated by the fact that scores are very highly skewed and restricted in variability. The standard error of measurement is small. The risk of making Type 1 or Type 2 classification error is small, considering that most candidates score very high on this examination.

Prosthodontics Test

As noted previously in this technical report, this test consists of three procedures: (1) cast gold crown, (2) porcelain-fused-to-metal crown preparation, and (3) ceramic crown preparation. The total score is 100 points. The cut score for pass/fail is 75.

Structure of the Data

A principal components factor analysis with varimax rotation showed that the three subtests of the prosthodontics test represent a single factor. That is, the three procedures are highly related. Thus, an ordinary coefficient alpha is sufficient for estimating reliability.

Examiner Consistency

As reported in Ray and Cob (2017b), Table 18 lists the percent of agreement among the three examiners for all observations. As noted there, the degree of examiner consistency is high.

Table 18: Examiner Consistency (Percent) for Three Tasks

Procedures	Percent Agree	Items	Points
Porcelain-Fused-to-Metal Crown Preparation	82.6%	10	40
Cast Gold Crown	84.8%	10	40
Ceramic Crown Preparation	84.0%	11	44

Reliability

As noted previously, the estimation of reliability is a bridge to understanding what risks candidates have when their scores are near the cut score of 75. The estimation of reliability of a combination of scores is based on the reliability estimates of each sub-test and the sub-score total score variances. Table 19 provides conventional alpha reliability estimates. These estimates appear low, but this observation is mitigated by the fact that these scores are very high and negatively skewed. These conditions result in lower reliability estimates, but, more important, if the standard error of measurement is small, then the precision of pass/fail decisions can be very high.

As noted for other tests in this examination program, stratified alpha is used in instances where the subtests have a moderate to high degree of independence. Correlations among the three tests are 0.63, 0.55, and 0.55. This degree of independence among the subtests points to the use of stratified alpha coefficient (Haertel, 2006, pp. 77-8). Stratified alpha for these three subtests is 0.84, whereas coefficient alpha is 0.80. Thus, stratified alpha is superior in accuracy.

The standard error of measurement is 2.7. The zone of uncertainty ranges from 73 to 77. Only four candidates have a score in this range. Note that these candidates have very low scores

when compared with the majority of candidates. The median score for these 781 candidates is 93.6 and the distribution is negatively skewed (-5.49).

Table 19: Reliability for Three Tasks						
Sub-tests	N	Ni	Mean	S. D.	Skew	Rel.
Porcelain-Fused-to-Metal Crown Prep.	781	10	93	6.4	-5.1	0.68
Cast Gold Crown	781	10	91	6.9	-3.8	0.64
Ceramic Crown Preparation	781	11	92	8.1	-3.2	0.65
Total Score (including those with three scores and no zeros)	781	31	92	6.1	-5.5	0.84

Claim Supporting Validity

The three parts of this test are highly related, as they should be. Examiner consistency is very high. Performance is very high. Reliability is very high despite the extreme skew of the data. With the margin of error, only four candidates had a score that may lead to misclassification.

Periodontics

Structure of the Data

This test has four subtests of varying point totals. Scaling/Subgingival Calculus Removal is the most dominant of these involving 60 of the 100 points available in a candidate score. Correlations among the scores for the four subtests are very low. These coefficients range from 0.20 to 0.75. A principal components factor analysis with varimax rotation yielded a single factor with an eigenvalue of 2.31. All four subtests loaded on the factor with coefficients ranging from 0.61 to 0.88. Periodontics appears to be a single factor and not a set of disassociated skills.

Examiner Consistency

As noted in the Table 20 below, examiner consistency was very high (Ray & Cobb, 2017c).

Table 20: Examiner Consistency (Percent) for Tasks

Tasks	Percent Confirmed	Items	Total Points
Extra\Intraoral Assessment	92.3%	8	16
Periodontal Measurements/Gingival Recession	93.8%	16	12
Scaling/Subgingival Calculus Removal	82.8%	12	60
Supragingival Deposit Removal	97.1%	12	12
Total		48	100

Reliability

Table 21 reports descriptive statistics and coefficient alpha estimate of reliability. Alpha is 0.82. The standard error of measurement is 2.85. Scores ranging from 73 to 77 provide this zone of uncertainty. Practically, those candidates with scores of 72 to 78 fall in this zone.

Table 21: Descriptive Statistics and Reliability Estimates for Five Tasks and Total Score						
	N	Ni	Mean	S. D.	Skew	Alpha
Extra\Intraoral Assessment	690	8	15.6	1.9	-6.8	0.86
Periodontal Measurements/Gingival Recession	690	16	11.9	0.8	-10.7	0.86
Scaling/Subgingival Calculus Removal	690	12	56.6	6.8	-11.6	0.79
Supragingival Deposit Removal	690	12	11.8	0.8	-3.6	0.72
Total	690	48	95.9	8.3	-5.7	0.84

N is Number of Examiners; Ni is number of items, S. D. is standard deviation.

The stratified alpha reliability estimate for the total score is 0.84. The standard error of measurement is 3.3. The zone of uncertainty stretches from 72 to 78. Twelve test scores were found in this zone of uncertainty. However, these examinees were among the lowest in the distribution of scores, where the mean was 95.9 and the median was a perfect 100.

Claim Supporting Validity

Examiner consistency is very high. The margin of error is in the acceptable range given the facts that the test scores are negatively skewed and the four subtests are independent of one another. The 12 candidates observed in this zone of uncertainty may be misclassified due to random error but their scores were substantially below the majority of examinees.

Restorative

Structure of the Data

There are four subtests with each subtest consisting of two parts: preparation and finish. A study of structure is possible with only two of the four subtests due to sample size limitations. Anterior Composite Preparation and Anterior Composition and Finish was the first subtest, and Posterior Composite Preparation and Posterior Composite Finish was the second subtest. A principal components factor analysis with varimax rotation revealed a single factor with strong loadings for each part of each subtest. The eigenvalue was 2.61, and about 65% of all variance was accounted by this factor. This result suggests that these two subtests represent a single factor–restorative proficiency.

Examiner Consistency–Amalgam Preparation and Finish

Table 22 presents information about examiner consistency from Ray and Cobb (2017d). Examiner consistency ranged from 79.2% to 83.5% for the four subtests, each containing two parts. As noted previously, a one-point examiner difference between two examiners is not materially harmful to candidates as the median is used instead of the mean. These levels of agreement are very high. In some instances, a criterion has been split into two different parts to facilitate examiners' specifying whether there is excess or a deficiency, overcut/undercut, or some other problem. Points are reassigned to accommodate that event.

Table 22: Examiner Consistency (% Agree) and Other Descriptive Statistics

Subtests	% Agree	Items	Total Points
1a. Anterior Composite Preparation	89.1	12	48
1b. Anterior Composite Finish	90.8	8	28
2a. Posterior Composite Preparation	90.4	11	44
2b. Posterior Composite Finish	88.5	8	28
3a. Amalgam Preparation	86.7	7	28
3b. Amalgam Finish	89.4	9	32
4a. Class II Slot Preparation	92.9	9	35
4b. Class II Slot Finish	88.5	8	28

Reliability

Because candidates took the first test and then chose from three other tests, we have several combinations of total scores. Due to the facts that average performance is very high resulting in a large negative skew and subscores have very low correlations, reliability

coefficients for each subscore were very low. However, the margin of error is also very small. Table 23 provides descriptive statistics for the four subtests.

Table 23: Descriptive Statistics and Reliability Estimate for the Three Subtests					
Subtests	N	Mean	S. D.	Skew	Rel.
1. Anterior Composite Preparation and Finish	750	91.5	21.2	-3.6	0.83
2. Posterior Composite Preparation and Finish	675	91.6	19.7	-4.0	0.86
3. Amalgam Preparation and Finish	47	77.6	37.1	-1.6	0.93
4. Class II Slot Preparation and Finish	30	93.4	18.4	-4.8	0.94

The stratified alpha coefficient for the combination of the first and second subtest in Table 23 was 0.88. The standard error of measurement is 5.7. Scores between 69 and 81 fall in this range of uncertainty. Of the candidates with total scores involving these two subtests, 18 candidates had scores in this range.

Claim for Validity

As with the other three tests, performance by well-trained candidates is very high. Given the skewed distribution, reliability appears very low, but considering the standard error of measurement associated with this reliability, only no candidates were observed in this zone of uncertainty due to random error.

SUMMARY OF VALIDITY EVIDENCE

CRDTS has designed and improved an examination that meets national test *standards*. Moreover, the argument presented in this report and the evidence assembled supports the claim for the validity of interpreting a test score as a measure of clinical dental competency.

To summarize this evidence:

1. Validity is the sine qua non. The *standards* cited in this technical report address validity directly and are well linked to the development of *CRDTS' NDE*, its administration, scoring, and reporting.

2. A basis was given for using a conjunctive scoring model that comprises four independent tests. The practice analysis and resulting studies involving data support that decision.

3. Item development includes the creation of tasks and scoring protocols. As noted in documentation, these are reviewed annually and polished and fine-tuned.

4. Examiner consistency is very high and this fact contributes to reliability. The resulting standard error of measurement helps develop a zone of uncertainty around the cut score of 75. Very few candidates have scored in this zone. These candidates are usually at the bottom of the test score distribution.

5. Examination administration is standardized. Documents report that administrative procedures are reviewed annually for the purpose of polishing and fine-tuning.

6. Examiners are carefully selected, trained extensively, validated, monitored, and retrained if scoring is not consistently high.

7. Scoring is very systematic with high degree of quality control.

8. Scores are reported responsibly.

9. All validity evidence is well documented in this report or other documents cited in the appendix.

10. The CRDTS website provide abundant information about all aspects of this examination program.

References

American Association of Dental Examiners (2003). *Guidance for clinical licensure examinations in dentistry*. Chicago: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2015). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates

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

Haladyna, T. M. (2007). Roles and importance of validity studies in test development. In S. M. Downing and T. M. Haladyna (Eds.) *Handbook of test development* (pp. 739-760). Mahwah, NJ: 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 .

Haladyna, T. M., & Rodriguez, M. R. (2014). *Developing and validating test items*. NY: Routledge.

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

Lane, S., Raymond, M., & Haladyna, T. M. (2016). *Handbook of Test Development*, 2nd ed. NY: Routledge.

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.

McCallin, R. (2016). Test administration. In S. Lane, M. Raymond, & T. M. Haladyna (Eds). *Handbook of test development,* 2nd ed. (pp.567-584). NY: Routledge.

Raymond, M. R. (2016). Job Analysis, practice analysis, and the content of credentialing examinations. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds). *Handbook of test development,* 2nd ed. (pp. 144-164). NY: Routledge.

Raymond, M. R., & 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.

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 (November 8-9, 2008). *CRDTS Dental Examination Review Committee* Report. Topeka, KS: Author.

CRDTS (April 17-18, 2009). *CRDTS Dental Examination Review Committee Report*. Topeka, KS: Author.

CRDTS (August 2009). *CRDTS Dental Examination Committee Meeting*. 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). *2010 New Examiners Orientation.* Topeka, KS: Author.

CRDTS (January 15-16, 2011). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.

CRDTS (April, 29-30, 2011). *Dental Examination Review Committee Meeting*. Topeka, KS: Author

CRDTS (August 25, 2011). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.

CRDTS (January 14-15, 2012). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (April 27-28, 2012). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (August 23, 2012). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (January 12-13, 2013). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (April 26-27, 2013). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (August 22, 2013). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (2013). *CRDTS Dental Examination Occupational Analysis*. Topeka, KS: Author.
CRDTS (January 11-12, 2014). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (August 22, 2014). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (2014a). *CRDTS' National Dental Examination Technical Report for the Year Ending 2014*. Topeka, KA: Author.
CRDTS (2014b). *New examiners orientation*. {Power Point Presentation}. Topeka, KS: Author.
CRDTS (2014c). *CRDTS' Security Measures*. Topeka, KS: Author.
CRDTS (January 17, 2015). *Dental Examination Review Committee Meeting*. Topeka, KS.
CRDTS (April 11, 2015). *Dental Examination Review Committee Meeting*. Topeka, KS.
CRDTS (January 9, 2016). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (April 30, 2016). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (2016). *Examination Review Committee Analysis of the 2016 Dental Examination*. Topeka, KS: Author.
CRDTS (January 2017a). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (January 2017b). *Dental Examination Review Committee Meeting Worklist*. Topeka, KS: Author.
CRDTS (June 3, 2017). *Dental Examination Review Committee Meeting*. Topeka, KS: Author.
CRDTS (2017a). *Chief Examiners Manual*. Topeka, KS: Author
CRDTS (2017b). *Dental Examination Candidate Manual*. Topeka, KS: Author.
CRDTS (2017c). *Dental Examiner's Manual*. Topeka, KS: Author.
CRDTS (2017d). *Examination Review Committee Analysis of the 2016 Dental Examination*. Topeka, KS: Author.
CRDTS (2017e). *Examiner Profile Statistics for the 2017 Clinical Manikin Examination: Endodontics and Fixed Prosthodontics*. Topeka, KS: Author.
CRDTS (2017f). *Periodontal Final Evaluation Form*. Topeka, KS: Author.
CRDTS (2017g). *Restoration Examination: Procedure*. Topeka, KS: Author.
Klein, S. P. (April 15, 2008). *Technical Report: Class of 2007–American Dental Licensing Examination*.
Klein, S. P., & Bolus, R. (May 9, 2008). *How many examiners are needed for case acceptance decisions?* Authors.

Klein, S. P. (November 2, 2010) *Technical Report: Class of 2009–American Dental Licensing Examinations*. Author.

Ray, L., & Cobb, K. (2017a). *Annual Report to Regional Schools of Dentistry. Dental Examination Results*. Topeka, KS: Author.

Ray, L. & Cobb, K. (2017b). *Examiner Profile Statistics for the 2017 Clinical Manikin Examination*: *Endodontics and Fixed Prosthodontics*. Topeka, KA: Author.

Ray, L., & Cobb, K. 2017c). *Examiner Profile Statistics for the2017 Clinical Periodontal Examination.* Topeka, KS: Author.

Ray, L., & Cobb, K. 2017d*). Examiner Profile Statistics for the2017 Clinical Restorative Examination*. Topeka, KS: Author.