



Central Regional Dental Testing Service, Inc.

An Evaluation of the
Central Regional Dental Testing Service's
National Dental Hygiene Examination

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

January 26, 2011

Acknowledgments

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

Dr. Thomas M. Haladyna
Phoenix, Arizona
February 2011

Table of Contents

Introduction	1
Part I: Purpose of This Evaluation	2
Part II: Description of the <i>National Dental Hygiene Examination</i>	3
Part III: Validity	5
Part IV: <i>Standards for Educational and Psychological Testing</i>	8
Part V: Legal Defensibility	11
Part VI: Validity Evidence	12
1. Content-related Validity Evidence	12
2. Item Quality	16
3. Reliability	19
4. Examination Administration	25
5. Recruitment, Training, Evaluation, and Retention of Scorers and Scoring	26
6. Scaling and Comparability	28
7. Standard Setting	29
8. Reporting	30
9. <i>Candidate's Manual</i> and Rights of Test Takers	31
10. Security	32
Part VII: Summative Evaluation	34
References	36
Appendix: Archive of Cited Documents Providing Validity Evidence	38

INTRODUCTION

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

CRDTS has engaged this author to conduct an independent evaluation of its *National Dental Hygiene Examination (NDHE)*. This author has conducted such evaluations previously and has written extensively on the subject.

This evaluation report has seven parts:

Part I addresses why this evaluation is being conducted.

Part II describes the *NDHE*.

Part III discusses validity and the investigative process known as *validation*.

Part IV identifies professional testing standards that apply to this test.

Part V briefly discusses the topic of legal defensibility.

Part VI reports the validity evidence collected to support validity.

Part VII is a summative evaluation.

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

PART I: PURPOSE OF THIS EVALUATION

Organizations like CRDTS provide test scores of candidates for licensure as dental hygienists for state boards. Although CRDTS and other regional testing services design, develop, and administer these tests to candidates for licensure, state boards have the responsibility for licensing these candidates. Thus, responsibility for validity belongs to each state and not the testing agency that develops, administers, and scores examinations.

Why has CRDTS engaged this evaluator to conduct an independent evaluation of the NDHE Program? Responsible testing agencies, like CRDTS, regularly engage in self-evaluation and external evaluation to assess its strengths and weaknesses and improve its testing program. Moreover, these evaluations inform the public about the high quality of the testing services provided to these state boards. Periodic evaluation of any testing program is an effective way to gain insight into the quality of the testing program and to learn about ways to improve the program. That is why testing experts have long recommended and endorsed the external evaluation of testing programs (Buckendahl & Plake, 2006; Downing & Haladyna, 1997; Madaus, 1992). Evaluation entails considerable study of documents, some data analysis, and discussion with representatives of CRDTS. In this report, many references are made to the testing literature that bears on this evaluation. Also cited are documents originating from both CRDTS and the test development organization known as the American Board of Dental Examiners (ADEX), with which CRDTS was formerly associated. Also data from CRDTS tests administered in 2010 were analyzed and reported here.

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

- 1. Defining of professional competence needed to practice safely,*
- 2. Development of an examination that validly measures professional competence, and*
- 3. Validation of the interpretation and uses of examination scores.*

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

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

Part II: Description of the *NDHE*

The *NDHE* is a clinical performance test for candidates for licensure. CRDTS uses a 100-point scale. A total score of 75 points or higher is recommended to state boards for a passing decision. The scoring model used is compensatory—that is, only a total score is important for making a pass/fail decision. Although the *NDHE* has five subscores, performance on any subscore is only one factor in determining a total score.

Performance Points

Each of the five subtests consists of six or more performance items that are dichotomously scored.

1. Oral Evaluation—14 points. Seven scorable items, two points are awarded for each intra/extra oral structure described correctly
2. Periodontal Probing—12 points. Twelve measurements are made for pocket depth. One point is assigned for each correct performance.
3. Scaling—56 points. Fourteen measurements are made. For each measurement, the candidate can earn a score of 0 or 4. Examiners must agree that surfaces are acceptably debrided of subgingival calculus.
4. Supragingival deposit removal—6 points. Six items are scored. One point is awarded for each item. Scoring is 0 or 1. Each tooth needs to be free of supragingival deposits for the candidate to earn a point.
5. Tissue management—12 points. Six items are scored. Each item is scored 0 or 2. To earn two points, the tissue surrounding each tooth must be free of damage and well managed.

Penalty Points Deductions

Penalty points can be deducted from the total score for various infractions. In all cases, examiners must form a consensus in their judgment. The penalty deductions are well explained in the candidate's manual (CRDTS, 2011b).

Patient does not qualify. The most serious infraction is a 7-point penalty for submitting an unqualified patient. If the second submitted patient is unqualified, another 7-point penalty is assessed. Subsequent submissions do not result in penalties. However, if a patient is not qualified, the candidate fails the test.

Critical issues in tissue management. A critical tissue trauma error will result in failure. These errors include the following: damage to four or more areas of gingival tissue, lips, or oral

mucosa, an amputated papillae, an exposure of the alveolar process, laceration or damage that requires suturing or perio packing, an unreported broken instrument tip found in the sulcus, or one or more ultrasonic burns requiring follow-up treatment.

Violations of standards. Penalty points are assessed for any violations of treatment standards: infection control, record keeping, patient management, professional conduct and demeanor and time penalties.

Professional demeanor. Penalty points can be assessed for infractions. Any infraction can be assessed more than once. The areas for possible penalty deductions include improper record keeping (2 points), failure to properly complete anesthetic documentation (2 points), professional demeanor (2 points), patient management (5 points), asepsis violation of disease barrier technique (2 points), gross violation of aseptic technique (10 points), time penalty (1-15 minutes late) 10 points, time penalty (16 or more minutes late)—dismissal from examination, unprofessional conduct—dismissal from examination

Scoring

All scoring is done by trained examiners. Examiners are highly qualified subject-matter experts (SMEs). Examiners receive extensive training in evaluating candidate performance on a patient. This training also includes calibration—where the precision and accuracy of their judgments is improved. At the end of an examination cycle, each examiner is subject to an evaluation of their precision and accuracy of their examiner performance. Feedback to examiners allows them to self-evaluate and improve. This information is also used to qualify examiners for future examinations.

Subsequent sections of this evaluation will deal with important aspects of test design, content, and examiner precision and accuracy.

PART III: VALIDITY

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

The *Standards for Educational and Psychological Testing* (American Educational Research Association-AERA, American Psychological Association-APA, & National Council on Measurement in Education-NCME, 1999) provides guidelines for evaluating validity. Additionally, the American Association of Dental Examiners-AADE (2005) issued guidelines for clinical performance examinations that include both dentistry and dental hygiene. These guidelines for validity were applied in this evaluation.

The questions for which we seek answers regarding validity are:

1. What does an examination score obtained from the *NDHE* mean?
2. How valid is it for a state to make a pass/fail decision based on this examination score?

Thus, validity does not address an examination, so the term *examination validity* or *test validity* is inappropriate. Validity focuses on the meaningfulness of an interpretation and the reasonableness of using the test score to pass or fail a candidate.

Validation

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

1. an argument that lays out what competence is to be measured and how CRDTS plans to measure it;
2. a claim is made that CRDTS' *NDHE* test scores are valid measures of the competence of a dental hygiene candidate, and that using a test score for making a pass is defensible/fail decision;
3. validity evidence is assembled and related to this argument and claim; and
4. a professional judgment that incorporates this argument, claim, and evidence into a summary judgment.

For a positive evaluation, the argument has to be sound and compelling, the claim just, and the preponderance of evidence supporting the claim. Negative evidence should be inconsequential. Negative evidence usually leads to recommendations to study, assess, and eliminate or reduce the factors causing this negative evidence. Validity studies are one remedy. By studying negative evidence and seeking remedies, validity is increased.

Table 1 shows the constituent elements in validation.

Table 1: Validation of CRDTS's <i>NDHE</i>	
Argument	The American Dental Association administers a <i>National Board Dental Hygiene Examination</i> . This examination measures the knowledge and skills thought to be necessary for safe and competent practice as a dental hygienist. This examination derives principally from a practice analysis of the profession of dental hygienists. The CRDTS <i>NDHE</i> is a clinical performance examination intended to measure competence of dental hygienists directly. These two examinations represent complementary aspects of the competence of dental hygienists. CRDTS's <i>NDHE</i> is the capstone in this licensing process for licensed dental hygienists.
Claim About Validity	CRDTS claims that candidate scores from its <i>NDHE</i> represent a level of competence of the dental hygienist candidate. The results of the test can be used with confidence by participating states, along with other criteria, to make licensing decisions for candidates who want to practice in that state.
Evidence Supporting the Argument	This evaluation report provides validity evidence of many types that are based on national test standards. CRDTS's documents cited in this report offer validity evidence supporting this argument.
Evidence Weakening the Argument	In this report, to the extent possible, evidence is displayed that weakens this argument. In the judgment of this evaluator, this kind of evidence as discussed in this report is inconsequential to validity. Nonetheless, CRDTS should consider threats to validity and act accordingly to diminish each threat. By that, CRDTS strengthens the evidence supporting the argument and the claim for validity.
Lack of Evidence	If evidence is missing, it is the responsibility of CRDTS to gather such evidence in the future as it contributes to increasing validity.

Summative Judgment	This evaluator considers the argument, claim, and evidence before making a judgement about validity of CRDTS scores as (1) a measure of clinical competence, and (2) for use by participating states in making pass/fail decisions.
--------------------	---

Validity Evidence Used in This Evaluation

The body of evidence to be presented is listed in the Table of Contents. This evidence should be evaluated holistically. Part VI of this report presents this validity evidence. This evidence includes recommended procedures, documentation, and statistical analyses. This evidence is used in the same manner that a jury weighs evidence and decides what supports either the prosecutor’s claim or the defense’s claim.

Evidence Weakening the Argument

No examination reaches its ultimate in validity. All examinations undergo improvements in validity in an evolutionary path. In any evaluation, the evaluator is responsible for truthfully exposing threats to validity (Cronbach, 1988). According to Messick (1989), two kinds of evidence that weaken validity are construct under-representation (CUR) and construct-irrelevant variance (CIV). Construct is another name for the domain of tasks that comprise dental hygiene competence. This part of the evaluation seeks to uncover evidence that may undermine validity. Naturally, CRDTS and its client states do not want such evidence to be strong, but its detection and eventual treatment are important steps in strengthening the overall validity argument and related claim for validity. Every examination is only as strong as its weakest link.

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

Another way of considering fidelity is to begin with the target domain of tasks that a dental hygienist performs. The target domain is the ideal test. The clinical performance test should contain items (tasks) that are identical or very similar to actual practice. The essential question is: Does our examination resemble those tasks that a dental hygienist performs often and best represents the profession? The test is a sample of the target domain, but the actual items are prioritized in terms of how often the task is performed in the profession and how critical the task is to professional practice.

PART IV: STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING

The *Standards for Educational and Psychological Testing* was published in 1999 by the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME). A large, representative committee of testing experts and other qualified volunteers participated in developing these guidelines. For this evaluation, these guidelines are used and often cited. All of the referenced guidelines bear on the overall judgment of validity. A set of new standards is being developed, but these new standards will not be published until 2012 or later. That is why the current standards are used for this evaluation. The American Association of Dental Examiners (2005) published *Guidance for clinical licensure examinations in dentistry*. These guidelines also apply to the examination in dental hygiene. Although not specifically cited, these guidelines also apply to this evaluation. The two sets of guidelines are very similar in terms of principles related to validity.

Table 2 on the next page summarizes some more important standards used in this document. Of the many categories that appear in that table and throughout this report, several notable omissions exist that deserve special treatment here.

Chapter 6: Documentation. This evaluation report contains *all* documentation made available by CRDTS used for the validity claim stated in this evaluation. This chapter has many categories of validity evidence. This report is one type of documentation. CRDTS keeps an archive of documents that bear on validity. Chapter 6 should be used as a guide for documenting its validity evidence. This documentation should be viewed as a kind of insurance that can be used to defend against criticism, legal challenges, and inquiries about the quality of CRDTS's examinations. Other information about the importance of documentation includes Becker and Pomplum (2006) and Haladyna (2002).

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

Chapter 9: Linguistic background. As this clinical performance examination involves patient treatment under simulated natural conditions involving patient-hygienist interaction, no threat due to inadequate linguistic background is perceived. Most of the candidates are trained in the United States and received their degree from a dental hygiene school. Foreign trained candidates often have difficulty with the English language. These candidates should also be treated fairly. While most candidates seeking licensure in the United States have adequate skill in

the English language, having linguistic or hearing difficulties sufficient to require an interpreter is not uncommon for patients. CRDTS permits interpreters to be available as needed outside the candidate’s operatory or the examiner station. All examination sponsors should always be alert to any threat arising from a lack of understanding of the recommended procedures for this examination or other factors that may jeopardize a candidate whose primary language is not English. A subtle point is that language should be appropriate for the practitioner. This examination should not simplify the language to accommodate an English language learner, because part of the professional responsibility in licensure is to ensure that the licensee has sufficient verbal ability to read, write, speak, and listen in English at an appropriate level for the profession of dental hygienist.

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

Chapter 10: Testing individuals with disabilities. Page 3 of the 2011 *Dental Hygiene Candidate’s Manual* (CRDTS, 2010b) discusses provisions for testing candidates with disabilities. A key issue with CRDTS’s candidates who apply for disability is that each case is individually assessed regarding disability and then any accommodation in the administration of the test is done in a way that does not alter the competence being measured.

Chapter 11. The responsibilities of test users. This category of standards applies to CRDTS's participating states who use examination information. Overall, states should have access to all information bearing on the validity of using examination scores for making pass/fail decisions. This is a state's responsibility; it is not CRDTS's responsibility. However, CRDTS is responsible for giving all participating states such information that supports their uses of examination scores. CRDTS's *Dental Hygiene Candidate's Manual* (CRDTS, 2011b) is published every year and provides much information. CRDTS's website also provides public access to documents.

PART V: LEGAL DEFENSIBILITY

Besides providing the highest quality examination possible, CRDTS does not want to be challenged legally for adverse test score decisions that might be considered invalid. Such challenges are expensive to defend and if successful may lead to loss of credibility that can ultimately weaken and destroy an examination program.

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

In all circumstances, any examining board should have continued legal counsel that examines threats that arise from legal actions and its position in thwarting these threats. By engaging in this evaluation where validity evidence is collected and organized, CRDTS very effectively reduces the threat of legal action. Mehrens and Popham (1992) provide a useful discussion of legal threats and validity. An assertive and positive program of evaluation and examination program improvement is the best remedy for avoiding legal challenges.

PART VI: VALIDITY EVIDENCE

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

1. Content-related Validity Evidence

A very important type of validity evidence for a credentialing examination is content-related (Kane, 2006). The content of any professional competency test should also be informed by current practices, research, and advances in technology (Cobban, 2008). The content should be consistent with historical and current advances in standards (Cortell, 2008). CRDTS appears to have a consistent record of updating content of its *NDHE* (CRDTS, 2008a; 2008b). Not only has CRDTS continued to update content and address national standards for content, but it has engaged and responded to test experts' construct criticism for improving content and overall validity (Haladyna, 2009, May 6, 2009; Klein, May 11, 2008; Littlefield, March 18, 2009, November 15, 2009; Littlefield & Wallace, January 30, 2008).

The focus of content-related validity evidence as discussed in the *Standards* (AERA, et al., 1999, p. 156) can be summarized in this way:

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

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

A dental hygiene clinical examination should be based on a domain of tasks that are performed by competent dental hygienists. As stated previously, in validity theory, the collection of tasks is the *target domain*. Ideally, these tasks in the target domain are organized by important content topic descriptors and prioritized according to relevance to the profession and how frequently the tasks are performed in regular professional practice. The means by which the testing agency obtains information about the criticality and frequency-of-use of these tasks is a survey known as a *practice analysis*, which has been previously mentioned in this report.

Practice Analysis

When a member of ADEX, CRDTS benefitted from a practice analysis that provided the basis for the content of the current examination (Klein, April 27, 2008). Another document that provides information about this important survey is a letter from CRDTS law firm (dated November 7, 2008), which reports some results of this service in two attachments. The Buros Institute for Assessment, Consultation, and Outreach at the University of Nebraska was contracted to conduct this survey. More than 1,500 dental hygienists comprising a national sample completed the survey. Tasks were rated for frequency of practice and criticality. The tasks identified as most frequently used and most critical to the profession of dental hygiene were cross walked with current, extant testing programs. CRDTS testing program currently uses these recommendations in the design of the current examination program.

Fidelity

As fidelity is a critical feature of any credentialing test, the current CRDTS *NDHE* has five subscores representing five tasks that are regularly performed by dental hygienists. These tasks are also highly rated for their criticality to the profession. The target domain of actual tasks and the universe of generalization (tasks on the *NDHE*) appear to be congeneric. Thus, fidelity of the currently examination appears to be extremely high. However, the judgment of SMEs is the most effective means for this judgment.

Structural Evidence

Knowing the structure of the test data has a significant influence on the assessment of content-related validity evidence and also reliability. In this section, several studies are reported bearing on the structure of the data.

Ideally, we would like all observations on candidate performance to be highly interrelated (internally consistent). If so, a coefficient alpha is a testing industry standard for estimating reliability because the professional competence has a singular dimension. Coefficient alpha is intended for test data that is highly internally consistent (unidimensional). However, the practice analysis led to the establishment of five independent performances and five subscores. It is very likely that these five subscores are statistically and conceptually independent. Table 3 presents descriptive statistics for the five subscores of the examination and the penalty points assessed by

a separate scoring system.

Table 3: Descriptive Statistics for the Five Scales of the NDHE and the penalty points.

Scale	Points	Mean	S. D.	Skewness
1. Extra/intra oral assessment	14	12.5	2.6	-3.2
2. Periodontal probing	12	11.4	2.1	-4.8
3. Scaling	56	45.8	12.9	-1.9
4. Supragingival Deposit Removal	6	5.7	1.0	-5.0
5. Tissue Management	12	11.6	2.0	-5.2
Penalty	17	1.7	4.0	-2.3
Total Score	100	85.7	18.6	-3.0

Correlations among these scales of the NDHE are shown in Table 4. All correlations are positive and statistically and practically significant. These correlation coefficients range from 0.37 to 0.92. Excluding penalty points, which is not a content category, the median correlation is 0.78. There is some separation. A useful technique for assessing the dimensionality of these scores is factor analysis.

Table 4: Correlations Among Scoring Variables

Scale	1	2	3	4	5	6
1. Extra/intra oral assessment	/	/	/	/	/	/
2. Periodontal probing	0.78	/	/	/	/	/
3. Scaling	0.52	0.60	/	/	/	/
4. Supragingival Deposit Removal	0.78	0.91	0.61	/	/	/
5. Tissue Management	0.78	0.92	0.59	0.92	/	/
6. Penalty	0.43	0.50	0.37	0.51	0.51	/

Factor Structure and Factor Analysis

Is dental hygiene competence a singular ability or does it consist of five fairly independent abilities. The correlation results suggest that a moderate amount of codependency exists suggesting that there is a common factor but the five parts of the examination may be distinguishable.

A statistical procedure known as confirmatory factor analysis was done. Technically, the procedure involves a hypothesis and appropriate probing analysis to ascertain if the hypothesis is correct. The hypothesis is that dental hygiene consists of five independent factors; the sum of which comprises dental hygiene.

A technical note is inserted here for readers desiring a specific description of the procedure: A principal components factor analysis with quartimax rotation was used with a specification of five independent factors.

The principal component analysis with the quartimax rotation extracted a main factor with very high loadings (ranging from 0.46 to 0.55) that represented the third subscore (Scaling). This main factor had an eigenvalue of 3.78. The second factor was periodontal probing, which had factor loadings ranging from -0.43 to 0.61. The fact that some observations had low factor loadings was because performance levels on these observations were very high. Correlation and factor analysis are very sensitive to skewed data. Thus, these variables tend to be neutral for factor loadings. Tissue management was the third factor with factor loading ranging between 0.28 and 0.69. Factor 4 was not clear but had some elements of factor two. Note that this factor has very high performances thus discrimination of it as a viable factor is not possible with factor analysis. Factor 5 appeared to be a very weak but internally consistent measure of oral assessment.

Conclusion

First, CRDTS has provided a basis for the content of the *NDHE* examination as a result of using a survey of dental hygienists known as a practice analysis, which is highly recommended for all credentialing testing programs. Second, The practice analysis resulted in five subtests with unequal weighting. CRDTS has determined its weights via a consensus building exercise using its SMEs. This practice is widely used to establish scoring weights. Third, the correlation and factor analyses point to a set of five subscores that have some interdependency but enough separation to suggest that the five factors when combined do not form a single unidimensional measure of dental hygiene competence. As a step to increase confidence in this conclusion, coefficient alpha for the combination of these five measures was computed. It is a very low 0.57. This coefficient is an underestimation of reliability because the five scales providing a total score have some separation (independence). This conclusion has considerable implications for estimating reliability. These findings also support the rationale for at least five sections of the current *NDHE*.

2. Item Quality

The tasks for the current NDHE were the product of many years of development and refinement. CRDTS has recorded regular meetings of its Examination Review Committee that trace some of these refinements (CRDTS, August 26, 2006; August 24, 2007, August 22, 2008a, August 22, 2008b; August 22, 2008c, July 11-12, 2009, January 2010, July 10-11, 2010.

This section will address aspects of item quality of a statistical nature. Test items (observed performances) should be based on the practice analysis and resemble those tasks performed by dental hygienists in their professional practice. Once the items are developed and refined, we want to know how the items perform in actual testing situations. Do the items discriminate between competent and incompetent professional practice?

1. Extra/Intra Oral Assessment

The extra/intra oral assessment is a task performed by all dental hygienists. The Examination Review Committee weighted this measure 14 out of 100 points. Each of the seven scorable tasks is scored 2 or 0. The score is the consensus of three examiners. Average performance on these seven items is 89.3% (12.5 out of 14 points). Table 5 shows the performance of each observation (item) based on three examiners. All seven items have a negative skew, suggesting a high performance. Most scores are 2s. For example only 30 of 1322 scores for item two were zero. Correlations among these seven items ranges from -0.026 to 0.094. It would appear that each item is an independent observation.

Table 5: Mean and Standard Deviation (S. D.) for the Seven Oral Assessment Items

	1	2	3	4	5	6	7
Mean	1.85	1.96	1.91	1.71	1.84	1.81	1.82
S. D.	0.53	0.30	0.42	0.71	0.54	0.59	0.57

2. Periodontal Probing

This subtest consists of 12 measurements on two teeth. Periodontal probing is a task that dental hygienists perform in their professional practice. The weighting, which is 12% of the examination total, was based on a consensus of SMEs, as is appropriate. Each item is scored 0-1 again based on examiner consensus. The mean performance of each of these 12 items ranges from 0.961 to 0.999. There is virtually no variability to these items. Total scores on these 12-point scales range from 7 to 12 with a very negative skew. In fact, about 84% have perfect scores, and another 14.4% have scores of 10 or 11. The items are not interrelated; each item is an independent observation.

3. Scaling

Scaling involves 14 observations where scoring is 4 or 0. Performance on scaling varies as shown in Table 6. All these means were based on same sample of 1,332 candidates. As we can see, the means vary from 3.23 to 3.51. The first item had the highest mean and last item had the lowest mean. Correlations among these 14 items are positive and range from 0.126 to 0.278. These correlations are higher than observed with the other subscales. Performance on scaling items is lower when compared with these other subscales. Also, more variability in the means of the items was observed. Finally, there is more discrimination among high and low performances. The average score for these 1,332 candidates was 40.6 out of 56 points or 72.5%.

Table 6: Mean and Standard Deviation (S.D.) for 14 Scaling Items

	1	2	3	4	5	6	7
Mean	3.51	3.38	3.30	3.42	3.38	3.39	3.35
S. D.	1.32	1.45	1.43	1.41	1.45	1.48	1.48

	8	9	10	11	12	13	14
Mean	3.40	3.39	3.33	3.33	3.31	3.29	3.23
S. D.	1.42	1.44	1.49	1.49	1.51	1.53	1.57

4. Supraginval Deposit Removal

Table 7 presents descriptive statistics. The six items supply 6 of the 100 total points. Each item scores 0 or 1. The nearly perfect performance generates very little variance. Thus, this subscore contributes very little to reliability as there is very little differential information in this subscore. By this account, candidates would pass this subscale if the cut score for this subtest were 75%. The overall mean is 95.7%. More than 96% of these candidates scores 83% or higher. Correlations among these six items are very low. These coefficients range from -0.013 to 0.141. The two highest coefficients were for adjacent pairs (1 and 2) and (4 and 5). These items appear to be very independent of one another.

Table 7: Mean and Standard Deviation (S. D.) For the Six Items in Periodontal Probing

	1	2	3	4	5	6
Mean	0.99	0.98	0.99	0.98	0.99	0.98
S. D.	0.12	0.12	0.10	0.14	0.11	0.13

5. Tissue Management

Table 8 presents descriptive for the six items for tissue management. Each item is scored 0 or 2. The means of each item is very high. The total score mean is 99.2% of points possible. As with other subscales, candidates are expected to perform perfectly, and they do for the most part. Of the 1,332 candidates, three had a score of six and eight had a score of eight. So the test did identify a few candidates with deficient performance. However, because the test only provides 12 out of 100 points, low performance in this subscale cannot lower overall candidate performance much.

Table 8: Mean and Standard Deviation (S. D.) for Tissue Management Items.

	1	2	3	4	5	6
Mean	1.95	1.99	1.98	1.98	1.98	1.98
S. D.	0.31	0.13	0.17	0.18	0.19	0.17

Conclusion

Items were developed that reflected the results of the practice analysis. The performance on these items is typically very high. Scaling is one exception. The scaling items are more numerous and the average performance is typically in the .80s, whereas the other subscales have item performance levels in the .90s.

The weighting for the five subscales is based on consensus judgments of SMEs that are members of the Examination Review Committee. Such weighting decisions are appropriately made by this committee. The number of items appears adequate for the purpose of each subscale. All items reflect high fidelity to the construct of professional competence in dental hygiene as attested by the practice analysis.

3. Reliability

Every test score has an unknown degree of random error and a true score. This error can be positive or negative and large or small. There is no way to discover how much error is in a test score or the true score (unless every task in the target domain is administered and perfectly scored). For a candidate whose test score is close to 75 (the cut score), we have a concern that a pass/fail decision might be incorrect due to random error.

We have two kinds of errors of classification for pass/fail decisions. Either the passing candidate receives a fail decision when the candidate's true score is passing (equal or above 75) or the candidate receives a passing decision when the candidate's true score is failing (below 75). We call these classification errors Type 1 and Type 2. Reliability affords us some insight into the risk of misclassifying candidates whose true scores are close to the cut score of 75. For candidates whose scores are well below or well above the cut score of 75, we have very little risk of misclassification.

The *NDHE* that has five fairly independent subtests that lead to a total score. The risk of misclassifying a candidate as passing when the true score is failing or misclassifying a candidate as failing when a true scorer is passing depends on how many candidates have scores close to 75 on the 100-point scale and the standard error of measurement. To obtain an adequate estimate of the standard error of measurement requires an accurate estimate of reliability.

1. CRDTS uses three examiners for each observation. This step ensures a degree of internal consistency in rating each item that is crucial in establishing reliability. Results of examiner consistency are reported in this section for each subtest.
2. CRDTS has many observations (test items) per subtest. Reliability benefits by having many observations.
3. CRDTS has special scoring rules for critical deficiencies. (*CRDTS, 2011 Candidate's Manual*, pp. 10-12). Penalty points are not included in the reliability analysis, because these points are determined by a separate scoring rule.

An appropriate technique for estimating reliability for the *NDHE* is stratified alpha (Haertel, 2006, pp. 76-78). Haertel asserts that when the test score is a linear combination of subtests, conventional reliability methods greatly underestimate reliability; whereas stratified alpha does not. The *NDHE* total score is a sum of performance on these five subtests.

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

Reliability is not an end; it is a means to an end. The objective of estimating reliability is to obtain an estimate of the margin of error around the cut score so that the Board can assess the risk for misclassifying candidates whose true scores are close to the cut score of 75. Once reliability is properly estimated, the degree of random error is estimated and used to study the number of candidates whose observed scores fall near the cut score. Hopefully, the margin of error is very low and the number of candidates whose scores fall into this margin near the cut score is small.

For each subscore, examiner consistency is reported and the reliability of the subscale is estimated. These reliability estimates are then used to compute a total score reliability. The total score reliability is used to estimate the standard error of measurement.

1. Extra/Intra Oral Assessment

In this section, examiner consistency and reliability are reported for this 7-item subscale.

Examiner Consistency. As all scoring is dichotomous, (0 or 2), consistency is evaluated when all three examiners agree. Table 9 presents the percent of perfect agreement among three examiners and alpha reliability estimates for each item. The alpha coefficients are appropriate for each item as examiners are supposed to provide internally consistent judgments. Given the few items and the restricted range in scoring (0 or 2) and the high performances of candidates, these alpha coefficients are very high.

Table 9: Examiner Consistency and Coefficient Alpha for Each Item Score

1	2	3	4	5	6	7
76.8%	93.0%	84.2%	65.0%	79.4%	75.6%	79.4%
0.49	0.58	0.14	0.50	0.46	0.44	0.60

Reliability. The next step in this process is to estimate a stratified alpha coefficient for the combination of seven independent items to form an extra/intra oral assessment subscore. The result was a coefficient of 0.423. This is predictably very low. As there is little variability in these subscale scores, a reliability coefficient will be low. The margin of error in these test scores is also very low.

2. Periodontal Probing

Table 10 presents the percent of examiner consistency and coefficient alpha for each of the 12 item scores.

Table 10: Examiner Consistency and Coefficient Alpha for Each Item Score

1	2	3	4	5	6	7	8	9	10	11	12
84.6%	84.6%	88.8%	80.2%	97.1%	87.0%	88.3%	98.2%	91.0%	83.0%	97.0%	87.9%
0.422	0.102	0.475	0.338	0.424	0.409	0.457	0.322	0.383	0.321	0.164	0.331

Examiner Consistency. For the 12 observations (items) for this subtest, rater consistency ranged from a low of 80.2% to 98.2% with the median being 87.9%. The corresponding alpha coefficients ranged considerably. For those items where rater agreement was low, alpha was low. Also, when examiner consistency reached its highest level, alpha was again low, due to the ceiling effect and restriction in range. Generally, low examiner consistency is the cause of low reliability.

Reliability. To estimate reliability for this subscore, as with the other subscores, stratified alpha was used. The result was 0.654. Although this coefficient may seem low, this coefficient's magnitude is greatly affected by skewness of the data and restriction of range. Both conditions were present in these data.

3. Scaling

The scaling subtests consist of 14 independent observations (items) which are scored 0-4. As with all other subscales, three examiners give their judgments. Results are presented in Table 11.

Examiner Consistency. As shown in the table, rater consistency is in the moderate range. These percentages range from a low of 67.6% to a high of 74.4%. The internal consistency estimates for the degree of consistency among raters is moderate.

Table 11: Examiner Consistency and Coefficient Alpha for Each Item Score

1	2	3	4	5	6	7
74.3%	70.0%	72.7%	73.1%	73.9%	74.4%	72.7%
0.61	0.61	0.64	0.64	0.67	0.67	0.67
8	9	10	11	12	13	14
74.4%	72.7%	72.5%	71.9%	71.2%	69.4%	67.6%
0.68	0.66	0.67	0.67	0.66	0.64	0.67

Reliability. The stratified alpha coefficient for this subscale was 0.904, which is very high. Several factors contributed to this result. First, these items have some dependency (internal consistency). Second, the scores are lower than other subscale scores, so there is more discrimination and variability. Third, we have 14 items, each observed by three examiners with moderate consistency. Had consistency been higher, this coefficient would be higher. As this coefficient is very high and has the greatest weight in the examination, it portends that the total score reliability might be as high.

4. Supragingival Deposit Removal

This subtest consists of six observations; each observation is scored 0 or 1. With three examiners scoring each item, the subtest has 18 data points. However, the items appear to be very independent. This subtest yields a total of six of the 100 score points in the examination.

Examiner Consistency. The degree of examiner consistency shown below in Table 12 is very high. This result is mainly due to the fact the performance levels for this subtest are very high (98.6). Only 11 scores (0.8%) were lower than 83%. This test has very little impact on the overall test score.

Table 12: Examiner Consistency and Coefficient Alpha for Each Item Score

1	2	3	4	5	6
90.1%	92.4%	91.6%	92.3%	91.9%	92.1%
0.304	0.376	0.244	0.467	0.297	0.424

Reliability. Stratified alpha is 0.475. This coefficient is very low for a performance test but the result is strongly influenced by the fact that performance of these candidates reaches the highest level with little range in scores. The data is very heavily skewed (-4.234). Scores are very restricted.

5. Tissue Management

This subtest consists of six observations by three examiners for a total of 18 data points. The scoring is 0 or 2 for each item. This subtest yields 12 points of the 100 score points.

Examiner Consistency. As shown in Table 13 below, the percentage of agreement among examiners is very high for all six observations. Alpha internal consistency coefficients are also reported in the table. These coefficients are very low, which is due to the fact that the performance levels of the candidates are very high and the range of scores is restricted.

Table 13: Examiner Consistency and Coefficient Alpha for Each Item Score

1	2	3	4	5	6
94.4%	95.1%	95.1%	95.8%	94.2%	95.2%
0.291	0.239	0.403	0.462	0.383	0.335

Reliability. The reliability of this subscale is 0.585. As with the other subscales, the high performance and restriction in the range of scores results in low coefficients. The high examiner consistency argues for a small degree of random error.

Reliability of the Total Score

Reliability has been estimated for each of the five subscales and is reported in the Table 14 below. To estimate the reliability of the total score, stratified alpha is used again.

Table 14: Summary of Subscales

Subscale–1318 candidates	Mean	Stan Dev	Strat. Alpha
1. Oral Assessment	12.9 (92.1%)	1.6	0.428
2. Periodontal Probing	11.8 (98.0%)	0.6	0.654
3. Scaling	47.1 (84.1%)	10.4	0.904
4. Supragingival Deposit Removal	5.9 (98.6%)	0.1	0.475
5. Tissue Management	11.9(99.2%)	0.3	0.585
Total	89.6 (89.6)	9.4	0.897

Standard Error of Measurement

The main benefit of estimating reliability is to estimate the margin of error around the cut score of 75 on the 100-point scale. By doing that, we can ascertain how many candidates have scores in that zone of uncertainty. Once known, we can evaluate whether reliability is sufficient or not, and the subscale data provides clues about how to improve reliability if needed.

Based on the stratified reliability estimate of 0.897, the standard error is 3.02. We can construct a zone of uncertainty about the cut score of one standard error. That zone includes scores 71.98 to 78.02. We find 74 (5.6%) of the candidates with scores in that region of the 100-point test score scale. Those 79 and higher are more likely to have true passing score, and those with scores 71 and lower are more likely to have true failing scores.

Analyzing the conditions under which these results occurred, several observations are seemed salient.

1. Low stratified alpha estimates for four subscores above is not a major contributing factor, because performances on the oral assessment, periodontal probing, supragingival deposit removal and tissue management subscales are very high scores relative to the cut score of 75. As a result, these four subscales contribute very little error variance to the total score.
2. Scaling has the most weight in the examination, the highest reliability, and the lowest average performance. The highest degree of error variance comes from this subscale even though reliability is very high.
3. Whether a candidate passes or fails seems to depend on performance in scaling. The other four subscales have a tendency to compensate candidates for poor performance in scaling. For example, if scaling were conjunctively scored, 316 (24%) of all candidates would fail. Using the total score that number would be 137 (10%) failed. Thus, scaling has a very significant influence on whether a candidate passes or fails.
4. An argument used to counter any assertion that an examination has inadequate reliability is that most candidates score well above the cut score and are not in jeopardy of failing. Candidates whose scores are close to the cut score of 75 have performed in a marginal way and have thus placed themselves in jeopardy.

Conclusion

The current examination has good reliability. The margin of error around the cut score is small. There is very little that can be done to improve reliability. One way is to increase examiner consistency in scaling. Another remedy is to increase the number of observations for some subscales. Neither or both of these changes probably will have a large effect on reliability.

4. Examination Administration

This standardized examination has been slightly revised each year upon recommendation from the Examination Review Committee. One of the most informative and useful sources of information about the current examination administration is the *Hygiene Coordinator Notebook 2010* (CRDTS, 2010c). The book has seven sections containing coordinator materials, orientation materials, examiner calibration exercises, assistant materials, anesthesia materials, miscellaneous forms, and examination forms.

Another useful source of information about administration is the *2010 Dental Hygiene Examiner's Manual* (CRDTS, 2010b). This 52-page booklet provides information about the examiner's duties in the administration from training to follow-up after the examination. The *CRDTS 2011 Dental Hygiene Candidate's Manual* also provides specific information about administration that is suitable from the candidate's perspective. Regular meetings of the Examination Review Committee address many aspects of examination administration and document revisions in administration aimed at making it better (CRDTS, August 26, 2006; August 24, 2007, August 22, 2008, July 11-12, 2009; July 10-11, 2010). MacCallin (2006) provided a basis for the importance of examination administration as a type of validity evidence. Many features of the CRDTS examination administration are included in MacCallin's recommendations.

Conclusion

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

5. Recruitment, Training, Evaluation, and Retention of Examiners and Scoring

According to Klein (April 27, 2008), CRDTS has a well-established program for examiner training and calibration. A committee was formed for the purpose of establishing and maintaining an examiner preparation program. This committee defined the criteria for selection of examiners, reviews and monitors examiner reliability, assigns examiners to test sites, and selects chief examiners, coordinators, and team captains. Several documents attest to the well organized, efficient, and effective system for recruiting, training, evaluating, and retaining examiners. The first is the *Hygiene Coordinator Notebook–2010* (CRDTS, 2010c), and the second is the *2010 Dental Hygiene Examiner’s Manual* (CRDTS, 2010b).

Selection of Examiners

Eight factors are the criteria for examiner selection (CRDTS, 2010b). Examiners must be in good standing with their state board, have an active practice, possess good health, make a commitment to participate in two or three examinations, accept CRDTS standards and evaluation criteria, accept the training regimen, receive a nomination to serve, and must observe an examination if a new examiner. These criteria are consistent with the finest practices. All examiners must be SMEs.

Training and Evaluation of Examiners

Each examiner receives a copy of the most current *Dental Hygiene Examiner’s Manual* (CRDTS, 2010b). All examiners receive an orientation and undergo a calibration exercise to ensure that their judgments are accurate and consistent. New examiners receive additional orientation. Each year analysis is done to figure out the accuracy and consistency of examiners (Ray, August, 2010). This information is used to help examiners improve their accuracy and consistency. Also, examiners may not be retained for future examinations based on these analyses.

Scoring

The *Dental Hygiene Candidate Manual* (CRDTS, 2010b) provides the most recent update of the conditions for scoring including examiner ratings and penalty point assessments. All these decisions were reached by committee consensus and then approved by the Board.

Scoring is done on site and ratings are recorded electronically. 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 Hygiene Coordinators and Team Captains. Hygiene Coordinators contribute to the development of these exercises. Each year the exercises are reviewed, evaluated, and revised-if necessary. Also, the *Dental Hygiene Candidate Manual* is also revised.

After the examinations are administered, CRDTS annually produces reports of examiner performance, which are intended for examiner self-assessment (Ray, August 2010). These results are also used to evaluate examiners and to inform decision-making for future examiner assignments. A very useful feature of these reports is the presentation of graphs showing degrees of leniency and severity in examiner judging. Such information can be very useful in refining training and improving examiner consistency or, if justified, removing examiners who are 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, makes decisions regarding their effectiveness, looks for emerging leadership qualities as Team Captains or Hygiene Coordinators. They also review each examiner's Peer Evaluations, which are part of the profile reports. Every examiner is asked to evaluate their fellow team members at the close of each exam. These Peer Evaluations focus on the examiner's behavior, preparedness, adherence to protocol, and work ethic. The EEAC is empowered to change an examiner's assignment if they are not functioning well in a particular role, they may send letters to those examiners who are outliers in their profile reports, or terminate the examiner's assignments if their results or behavior is not appropriate.

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

Conclusion

The training of examiners by CRDTS is a highly refined activity that has received considerable attention over many years. No recommendations for improvement are offered.

6. Scaling & Comparability

A compensatory scoring model is used for the 100-point scale. The cut score is 75. As the test is standardized as to tasks (test items) and scoring, no scaling adjustments or equating is necessary. All tests cover the same content.

Because CRDTS keeps a record each year of mean scores for each subscale (CRDTS, 2011a), there is ample evidence that the mean scores are stable from year-to-year. The fluctuations in passing rates appear to be more associated with the professional competence of the sample. Thus, this record provides evidence that the scaling is comparable from test administration to test administration and year-by-year.

Two threats to validity appear possible. One threat comes from variability of examiner bias and inconsistency from one test site to another. However, these two problems are closely monitored (Ray, August 10, 2010). Another threat may arise from natural variability in patients and the problems they present to the candidate. This problem has been studied for more than 30 years. CRDTS considers this possibility and offers as a remedy patient acceptability criteria. Some of the specific steps taken to eliminate this threat are to require treatment selection of at least six and no more than 10 teeth, specifying the types of teeth in the treatment selection, and having the candidate present at least 14 surfaces for evaluation by the examiners.

Conclusion

Although the dental hygiene ability of candidates may vary from test site to test site, the fact that the test is the same for all candidates and examiner training and scoring is highly refined and accurate argues for comparable scaling from test site to test site.

7. Standard Setting

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

CRDTS monitors passing/failing rates by many categories as part of an effort to evaluate their standards (CRDTS, 2006, 2007, 2008a, 2009, 2010, 2011). This information provides an indication of the stability of these rates vis a vis the established cut score.

Conclusion

Because states mandate cut scores, testing agencies are placed in the uncomfortable position of needing a passing score study that produces a result that very likely is not 75 on this 100-point scale. One remedy is to conduct a passing score study and slide the scale so that the cut score (say the recommended cut score of 72.4) becomes 75 on the new scale. That way, the testing agency has fulfilled their responsibility in performing a passing score study, and also, they report the practical cut score at 75. In the jargon of testing, two scales exist: the raw score scale where the cut score is fixed, and the standard scale score where the cut score is 75. That way, both the testing agency and the state have satisfied their need for a standardized and valid cut score.

8. Reporting

Three kinds of score reports are distributed after an examination: (1) to candidates, (2) to jurisdictions, and (3) to dental hygiene schools.

Score Reports to Candidates

Two kinds of candidate score reports are sent. The report of a failing score is provided to the candidate with a justification/critique in the lower portion of the report. The report of a passing score simply provides the total score with a recommendation to a state to pass the candidate if other criteria for licensure have been met. Both score reports provide the total score.

Score Reports to Jurisdictions

For every examination site, a complete scoring report is prepared that presents scores for all candidates and other information.

Score Reports to Dental Hygiene Schools

A comprehensive score report is sent annually to each dental hygiene school (CRDTS, January 2010). This report provides an overview of the examination including its purposes. For every school the mean score on the examination is given along with the percent of candidates who passed and a quintile ranking. All subscores are reported in the same fashion.

Such scores can be validly used to identify strengths and weaknesses in each dental hygiene school's program and curriculum. However, it seems invalid to make comparisons among dental hygiene schools in terms of performance unless all candidates entering each program are of equal ability.

Conclusion

Score reports are responsibly and effectively designed. No recommendations are offered.

9. Candidate Manual and Rights of Test Takers

The *Dental Hygiene Candidate's Manual* is the official publication for candidates (CRDTS, 2011). This 56-page booklet contains many topics of interest and importance to candidates. An index at the end provides guidance for those wanting to find specific topics.

CRDTS website (<http://www.crdts.org/>) is also a source of information for candidates. The dental hygiene section of this website offers candidates application and eligibility requirements, examination calendar, the content of the examination and how it is scored, forms and manuals, online application, and orientations. The candidate orientation is taken online.

CRDTS has a complaint review process described on pages 7-8 in the *2011 Dental Hygiene Candidate's Manual* (CRDTS, 2011b). Candidates can voice a concern or complaint or formally petition for a review of the administration. Candidates are given a prompt hearing.

Failing Candidates

Failing scores are verified. A candidate who fails an 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 *2011 Dental Hygiene Candidate's Manual* (CRDTS, 2011b).

Conclusion

CRDTS provides a very effective candidate manual that is updated annually. Its website provides many services to candidates. All candidates are entitled to an online orientation. Candidates have an appeal process that is clearly described. Overall, CRDTS does an excellent job of respecting the rights of candidates as stated in our national testing standards (AERA, APA, & NCME, 1999).

10. Security

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

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

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

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

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

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

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

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

Because all treatment selections are checked by three examiners to ensure that the selection fulfills the criteria before a candidate is allowed to begin treatment and candidates are monitored during the examination, it is not possible for candidates to pretreat the teeth on which they will be tested or receive outside assistance with their procedures.

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

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

Conclusion

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

VII: SUMMATIVE EVALUATION

The body of evidence assembled in this report argues in favor of the claim that CRDTS provides highly valid score interpretations of dental hygienist clinical competence. The scores provided can be used with confidence for making pass fail decisions.

All testing programs can benefit from annual evaluation and short-term and long-range plans for continued improvement. CRDTS has benefitted from many years of test development and validation and its association with ADEX.

The following discussions are intended to elicit discussion for the purpose of short-term and long-range planning to continue to improve this already fine testing program.

Conjunctive Versus Compensatory Scoring

The *NDHE* test consists of five subscale with subtle yet different characteristics. Each subscale is well defined and justified in terms of dental hygiene competence. Evidence has been gathered and organized to support the uniqueness of each subscale. In the evaluation of the subscales and the entire test, four of the scales have very high performances that suggest that a conjunctive scoring condition exists. In other words, most candidates can pass any of these four subscales. However, scaling has a pronounced lower mean than the other subscales (CRDTS, 2011a). For the 2010 examination year, 23.4% would clearly fail a conjunctively scored scaling test. Thus, the compensatory model benefits these candidates by boosting overall performance by doing better on the other four subtests. CRDTS should examine the rationale for compensatory scoring and ascertain if this is what is desired. The conjunctive scoring model is preferred because it treats each subtest as a test with its own cut score (say 75%). If applied to these data, many more candidates would fail due to low performance on scaling. Perhaps a modified standard where conjunctive elements are added to the compensatory scoring model might make this tendency for candidates to compensate a low scaling score with high scores in the other four subtests.

Technical Reporting

CRDTS NDHE would benefit by having an annual technical report where validity evidence is gathered, organized, and reporting. The benefit of such reports are many. First the validity evidence reported in this document is annually issued to the public and CRDTS' constituency. Second, such reporting often identifies areas of concern for improvement, which increases validity. Third, the reports provide a historical basis for trends that may affect future revisions of the examination program.

Practice Analysis

The previous practice analysis has greatly affected the content of the examination. Practice analysis needs to be conducted periodically, and its results need to be used to make appropriate revisions in content or validate the current examination content. The practice analysis should be published. A new practice analysis should be done soon.

Cut Score

The current cut score set by states is legally defensible. However, the testing industry standards call for a cut score study to set a fair cut score. A white paper should be written by an expert that addresses the conflict between what state legislatures do and how testing agencies set cut scores. Clearly there is a disconnect that is not caused by the testing agency. However, all testing agencies need a policy and appropriate procedures for dealing with how the cut score is established.

Item and Test Specifications

The practice analysis should produce a document known as item and test specifications that shows the connection of the results of the practice analysis to the test design, including the specific items used.

Documentation

All meetings dealing with test development, analysis, and validation should be documented. Minutes should be published with dates included. All documents bearing on validity should be archived and referenced either in a technical report or an independent evaluation, as is evidenced in this appendix.

Closing

CRDTS deserves much credit for providing highly valid test scores to its constituency that can be used confidently for making pass/fail decisions for licensure purposes. CRDTS has a proven record of periodic review of its testing program and testing program improvement.

References

- American Association of Dental Examiners (2003). *Guidance for clinical licensure examinations in dentistry*. Chicago: Author.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Becker, D. F., & Pomplum, M. R. Technical reporting and documentation. In S. M. Downing and T. M. Haladyna (Eds). *Handbook of Test Development* (pp. 711-724). Mahwah, N.J.: Lawrence Erlbaum Associates.
- Buckendahl, C., & Plake, B. S. (2006). Evaluating tests. In S. M. Downing and T. M. Haladyna (Eds). *Handbook of Test Development*, pp. 725-738. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cronbach, L. J. (1988). Five perspectives of the validity argument (pp. 3-18). In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Downing, S. M., & Haladyna, T. M. (1996). Model for evaluating high-stakes testing programs: Why the fox should not guard the chicken coop. *Educational Measurement: Issues and Practice*, 15, 5-12.
- Downing, S. M. & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10,61-82.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.) *Educational Measurement*, 4th edition, pp. 65-110. Westport, CN: Praeger.
- Haladyna, T. M. (1998). *An evaluation of the Western Region Examination Board Dental Hygiene Examination*. Phoenix: Author
- Haladyna, T. M. (2002). Supporting documentation: Assuring more valid test score interpretations (pp. 89-108). In J. Tindal & T. M. Haladyna (Eds.) *Large scale assessment for all students*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Haladyna, T. M. (2009). *White paper on validity of the NDHE*. Phoenix: Author.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27.
- Haladyna, T. M., & Hess, R. K. (1999). Conjunctive and compensatory standard setting models in high-stakes testing. *Educational Assessment*, 6(2) 129-153 .
- Kane, M. T. (2006a). Content-related validity evidence. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development*, pp. 131-154. Mahwah, NJ: Lawrence Erlbaum Associates.
- Madaus, G. F. (1992). An independent auditing mechanism for testing. *Educational Measurement: Issues and Practice*, 11(1), 26-31.
- McCallin, R. (2006). Test administration. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development*, pp. 625-652. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mehrens, W. A., & Popham, W. J. (1992) How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education*, 5(3), 265-283.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan.

Raymond, M. & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing and T. M. Haladyna (Eds). *Handbook of Test Development*, pp. 181-224. Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix: Archive of Cited Documents Providing Validity Evidence

- ADEX. (2008). *ADEX Dental Hygiene Committee Report to CRDTS Dental Hygiene ERC*. Cobban, S. J. (July 2008). Commentary on “Moving research knowledge in dental hygiene practice”, *Access*, 46.
- Cortell, M. (July 2008). Understanding and implementing the 2008 American Dental Hygienists’ Association Standards for Clinical Dental Hygiene Practice. *Access*, 43-44.
- CRDTS (2006). *Analysis of the 2005 Dental Hygiene Examination*. Topeka, KS: Author.
- CRDTS (August 26, 2006). *CRDTS Dental Hygiene Examination Review Committee Committee Meeting*. Topeka, KS: Author.
- CRDTS (2007). *Analysis of the 2006 Dental Hygiene Examination*. Topeka, KS: Author.
- CRDTS (August 24, 2007). *CRDTS Dental Hygiene Examination Review Committee Committee Meeting*. Topeka, KS: Author.
- CRDTS (2008a). *Analysis of the 2007 Dental Hygiene Examination*. Topeka, KS: Author.
- CRDTS (2008b). *Background and current issues with the ADEX Dental Hygiene Examination*. Topeka, KS: Author.
- CRDTS (August 22, 2008a). *CRDTS Dental Hygiene Examination Review Committee Committee Meeting*. Topeka, KS: Author.
- CRDTS (August 22, 2008b). *Recommendations for the CRDTS Hygiene Examination Review Committee*. Topeka, KS: Author.
- CRDTS (August 22, 2008cc). *Attachment A & Attachment B*. Topeka, KS: Author.
- CRDTS (2009). *Analysis of the 2008 Dental Hygiene Examination*. Topeka, KS: Author.
- CRDTS (July 11-12, 2009). *CRDTS Dental Hygiene Examination Review Committee Committee Meeting*. Topeka, KS: Author.
- CRDTS (2010a). *Analysis of the 2009 Dental Hygiene Examination*. Topeka, KS: Author.
- CRDTS (2010b). *2010 Dental Hygiene Examiner’s Manual*. Topeka, KS: Author.
- CRDTS (2010c). *Hygiene Coordinator Notebook-2010*. Topeka, KA: Author.
- CRDTS (January 2010). *2010 Annual report to program directors on dental hygiene examination results*. Topeka, KS: Author.
- CRDTS (July 10-11, 2010). *CRDTS Dental Hygiene Examination Review Committee Committee Meeting*. Topeka, KS: Author.
- CRDTS (2011a). *Analysis of the 2010 Dental Hygiene Examination*. Topeka, KS: Author.
- CRDTS (2011b). *2011 Dental Hygiene Candidate’s Manual*. Topeka, KS: Author.
- Haladyna, T. M. (2009). *Observations and Opinion About American Dental Hygiene Examination*. Phoenix: Author.
- Haladyna, T. M. (May 6, 2009). *Comments on ADEX’s White Paper Dated March 18, 2009*. Phoenix: Author.
- Klein, S. P. (April 27, 2008). *Technical report on the ADEX/CRDTS Dental Hygiene Licensing Examination: Class of 2007*.
- Klein, S. P. (May 11, 2009). *Technical report on the ADEX/CRDTS Dental Hygiene Licensing Examination: Class of 2008*. Santa Monica, CA: Author.
- Littlefield, J. (April 25, 2009). *Comments on ADEX’s White Paper Dated March 18, 2009*. San Antonio, TX: Author.

- Littlefield, J. (November 15, 2005). Letter to credits. San Antonio, TX: Author.
- Littlefield, J., & Wallace, J. (January 30, 2008). *Psychometric recommendations for the American Dental Hygiene Licensure Examination*. San Antonio, TX: Authors.
- Ray, L. (August 2010). CRDTS Examiner Profile Service: Examiner Profile Statistics–2010. Topeka, KS: CRDTS.