

Technical Report for CSW Computer Simulations LLC

(Computer administered dental licensing testing developed by Central Regional Dental Testing Service (CRDTS), Southern Regional Testing Service (SRTA) and Western Regional Examining Board (WREB).

March 2010

Written by:

Del Hammond, M.A., Testing Specialist

Introduction

CSW tests were first administered in 2006 to candidates for dental licensure as part of the SRTA and WREB licensing tests. The tests were administered via the computer at Pearson VUE testing centers. The tests were designed to evaluate entry-level candidates' clinical application of knowledge and judgments necessary to provide periodontal and prosthodontic care. This report provides documentation describing test development and the post-test analyses from the first year of testing.

General History

Dentists from CRDTS, SRTA, and WREB (CSW) first met in November 2001 to develop a joint Prosthodontics test, reviewing models and questions that had been previously used independently by the three respective testing agencies.

A committee task force decided in February 2002 to pursue the concept of a computer-based test and met to evaluate presentations by three testing software developers in August of 2002. A contract for software development was awarded to Zoomorphix, an Australian firm with experience in developing performance test questions that rely on high quality graphics.

In November 2003 a joint CSW periodontal committee comprised of dentists and dental educators representing the three partnering organizations met to begin development of a computerized periodontal test that could be used to support licensure decisions.

Software and test development continued on both the periodontal and prosthodontic tests with the first tests administered via computer to candidates at Pearson VUE testing centers in March 2006.

Practice Analysis

An updated practice analysis and test specification was completed by the WREB prosthodontics subcommittee in 1999. The CSW prosthodontics committee comprised of dentists and dental educators from each of the three partnering organizations reviewed the WREB analysis and revised the test specification. The most current test specification for prosthodontics is dated October 27, 2005.

The CSW periodontal committee that is also comprised of dentists and dental educators from each of the three partnering organizations analyzed the results from the periodontal practice analysis and developed a test specification in November 2003. The current test specification for the periodontal test is dated July 10, 2004.

Item Development

The CSW prosthodontics and periodontal committees were composed of subject matter experts that included dentists who are university instructors in dental degree programs and practicing dentists who are dental examiners from the three testing agencies. The items used in the tests were required to be case specific in order to test clinical application skills. General knowledge questions were avoided as they would duplicate information that is gathered in other parts of the licensure testing process.

The prosthodontics committee used dental models that had previously been used in testing this discipline in dentistry by the three agencies. CSW's contractor, Zoomorphix, converted the models to

three dimensional computer models. The committee reviewed the previous items and item analyses. Items that functioned well (i.e. met CSW's criteria for acceptable item difficulty and discrimination) and were appropriate for the test specification were entered in the item bank. Additional items were developed by the subject matter experts to satisfy the requirements of the test specification. The item pool was intended to be large enough to support multiple forms of the test. Multiple forms are desired to protect item security and facilitate re-take opportunities for candidates who do not pass on their initial attempt.

The periodontal subject matter experts developed items based upon patient cases that were available at dental schools and private practices. The items for each form of the test cover medical history, oral history, assessment, prognosis, treatment plan, and re-evaluation on one patient case. As with the prosthodontic test, multiple cases were developed to protect item security and allow for re-take opportunities for candidates who fail on their initial attempt.

The test items for both tests were field tested with junior dental students at various dental schools using software that was similar to that which was used at the testing center. This method of item tryout did not duplicate the testing situation exactly, so the final decision on which items would be scored was postponed until after actual dental licensing candidates had completed the tests. Item functioning was then evaluated and some items were removed from the items to be scored before any decisions were made about candidates' performance. Subsequent revisions to items that were not scored have been field tested in unscored portions of the tests during candidate testing to gather information about the quality of the items prior to their operational use.

Both the periodontal and prosthodontics committees meet one or more times each year to add new items to the item bank and revise items based upon item analysis results and the currency of the items due to changes in dental practice.

Standard Setting and Equating

The Ebel method (Livingston & Zieky, 1982) of standard setting was employed using the committee subject matter experts as judges. This methodology is a systematic, test-based approach where committee members make independent judgments on each item of the test to provide a recommendation of the cut (passing) score necessary for the test based on its relative difficulty. The average recommendation across the panel is then passed along to the CSW Board for a final decision about the cut score for each of these tests. This method was used for the base reference form for Both the prosthodontics and periodontal tests.

Because the average difficulty of the items and the number of scored items for the different forms of each test were not precisely the same, a statistical strategy is used so that each form of the test is equal in terms of the meaning of the cut score. This is a strategy that insures fairness to candidates because it means that it does not matter which form a candidate takes, the meaning of the pass/fail decision is the same. Specifically, CSW uses equipercentile equating (Livingston, 2004; Kolen & Brennan, 1995) to provide equal treatment in the scoring of all candidates across forms of each respective examination. After data were available from initial testing, equipercentile equating was used, along with the Ebel study cut scores, to provide cut scores that resulted in an equal percentage of passing candidates for each form of the tests. After the initial testing year, equipercentile equating is used in subsequent years to provide cut scores that result in pass rates that are similar to the reference year. The equated scores are used to create tables to transform the raw scores into scores where 75 represents the passing standard.

Post-Test Analysis Values for Licensing Tests

Non-licensing tests are generally administered to a population of individuals whose abilities and knowledge vary among the test takers. Test scores for a test often follow a normal distribution and parameters such as the standard deviation are computed to describe the score distribution. The purpose of testing is often to determine the knowledge level of each of the individuals. One purpose of the tests, in educational settings, is to determine grades to be assigned to individuals. The items on the tests are those that test a broad range of abilities so that very high and very low performing individuals can be identified.

Licensing tests are used to identify individuals who meet the requirements for performing an activity or profession. There is no need to identify very high or very low performing individuals. These tests only need to separate the qualified individuals from the unqualified. The test items only need to be those that evaluate the minimum abilities required for the license. There is no need to find out who has knowledge or skill that is well above or below what is required by the license. Consequently, there is no need for items that are very difficult for qualified applicants. Most of the candidates are well qualified, so it is expected that most of them should be able to answer most of the questions.

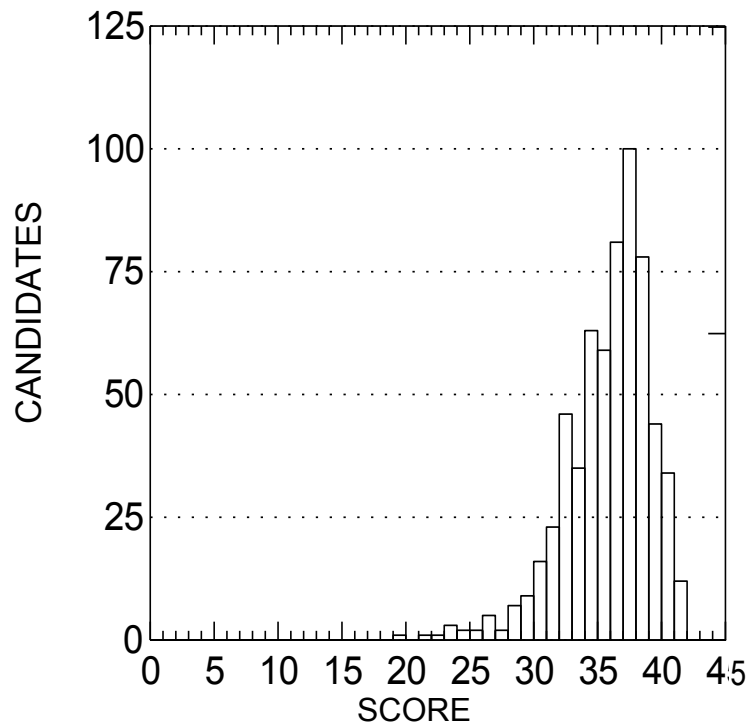
As a result, test items (questions) that would be difficult for someone who is not well prepared, may be easy for the well-prepared dental candidates who take the CSW tests. Consequently, many of the items will have a very low difficulty level on the post-test item analyses. These items are considered to be important by the subject matter experts on the item development committees. The distribution of these test scores would not be expected to follow a normal distribution, but will usually be leptokurtic and negatively skewed (peaked with most applicants scoring very high). Additionally, if test items are identified on post-test item analyses as being somewhat difficult (less than 50 or 60 percent of the candidates answer correctly), there is probably something wrong with those items. They may need to be revised or removed from the test. Those questions could be the result of item construction errors. Items that most candidates answer correctly will generally be retained because of the importance of retaining their content in the test, regardless of the lowering of the overall discrimination values and reliability estimates that will result.

These differences between general ability testing and licensing tests, regarding test question difficulty and differences in test taker knowledge and abilities, cause the formulas for reliability estimates to be of marginal value for evaluating licensing tests. The internal consistency reliability estimation formulas were designed for and are useful for educational and general ability testing. The basis for the formulas is the expected variability among candidate scores and the expected variability in the correctness of individual candidate responses. Those variabilities are very low for dental licensing testing. As a result, the formulas and “rules” for educational testing are not very useful for licensing tests. The formulas provide reliability estimates that underestimate (are lower than) true reliability.

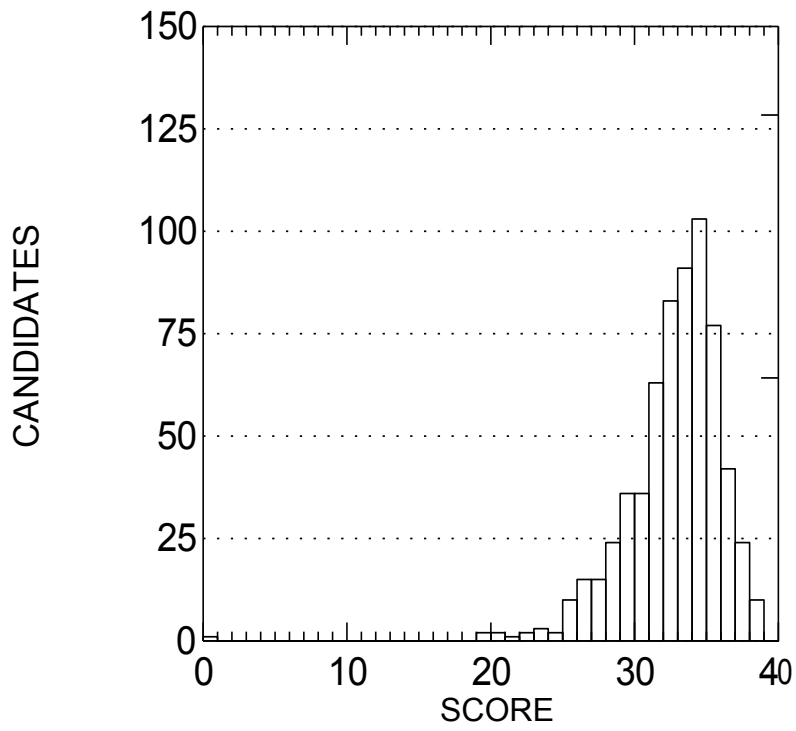
Results of Post-Test Analysis

For evaluating the characteristics of items and decisions on the test, CSW conducts a series of analyses. The results of these analyses are included as Table 1. First, we conduct internal consistency based reliability analyses to evaluate the extent to which the items are error free. These values range from 0.00-1.00 with values above 0.70 usually accepted as suggesting acceptable reliability for general testing. However, one of the assumptions of internal consistency based estimates, such as KR-20 (coefficient alpha), is that there is a range of ability in the underlying population of examinees. In dental licensure settings, we expect the abilities of candidates to be high

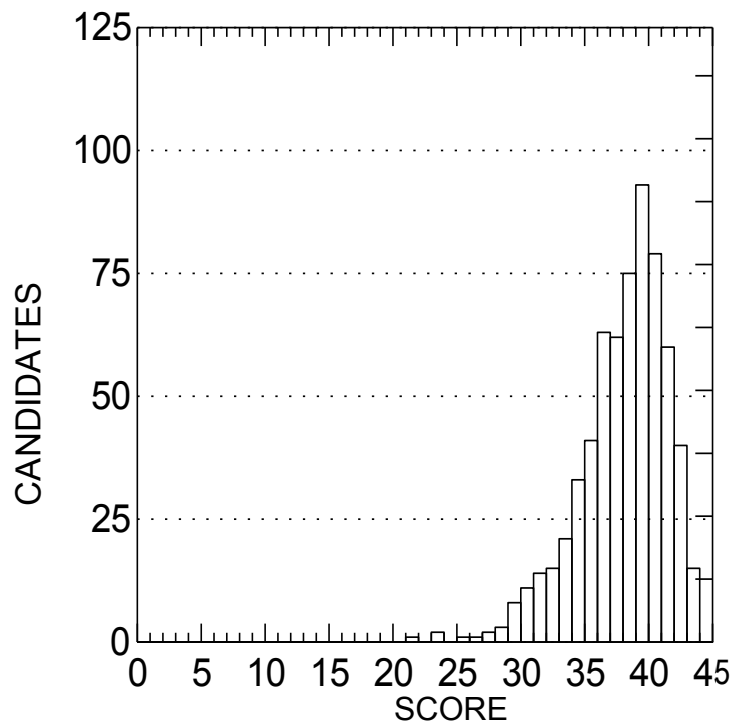
and more homogeneous than the abilities in a typical population of examinees in educational settings. In addition, the abilities evaluated on dental licensing testing are those that the test development committees have determined important and are also those that most entry-level dentists should be able to accomplish (may not be difficult for most candidates to answer, but are still considered important to ask). Consequently, if less than 60% of the candidates answer a test question correctly, then it is likely that there is some problem with the question. That question is removed from the test. Questions that candidates score very highly on (above 90% correct) are retained if the item analysis shows a positive discrimination effect for those questions. A highly qualified population of test takers being scored on abilities, that most should possess, results in high scoring with limited variability in candidate scores. The formulas for estimating reliability are based upon score variability. High scores and limited variability suppresses estimated reliability (and increases associated estimates of testing error) to a point that these estimates may have little practical value. The following histograms show the scores for the four periodontal and the four prosthodontic test forms that show the high scoring that was predicted.



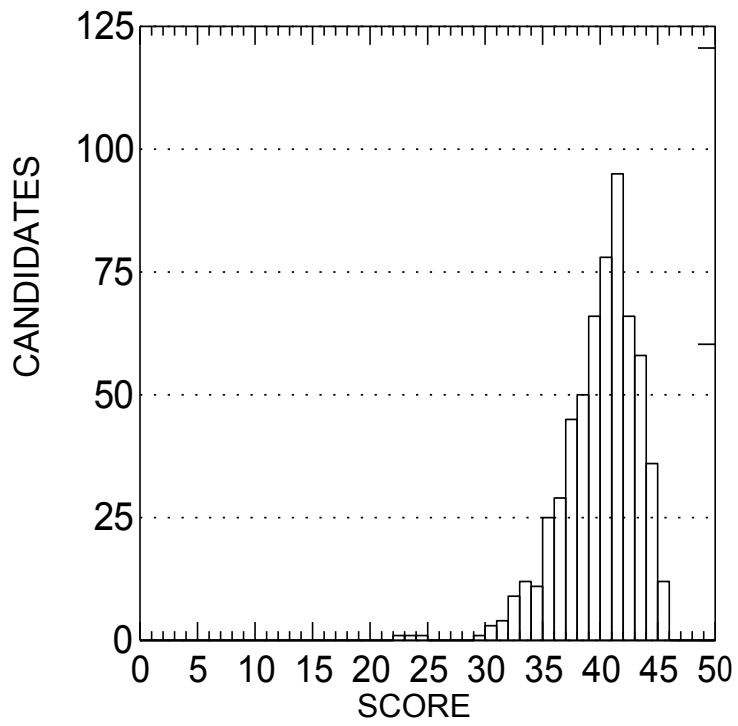
PERIODONTAL FORM 1



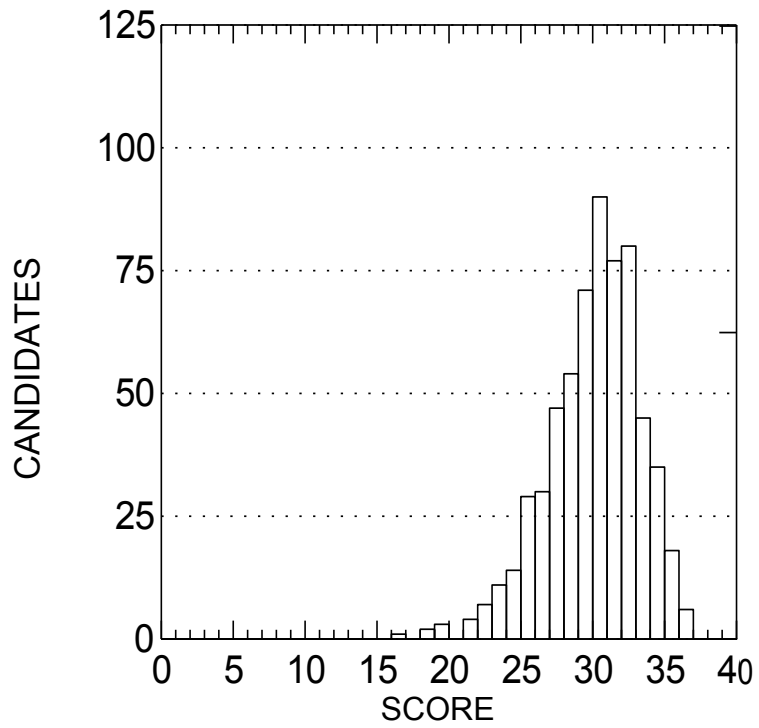
PERIODONTAL FORM 2



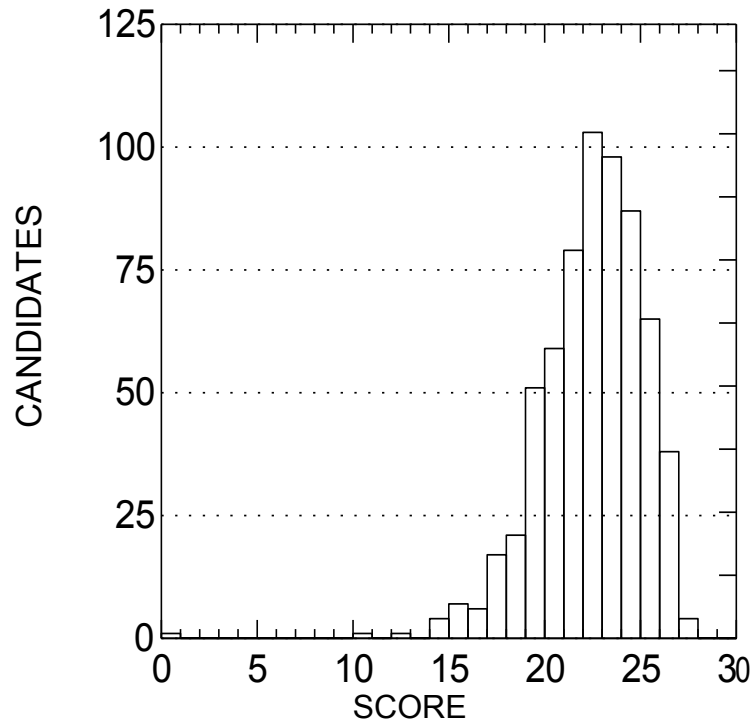
PERIODONTAL FORM 3



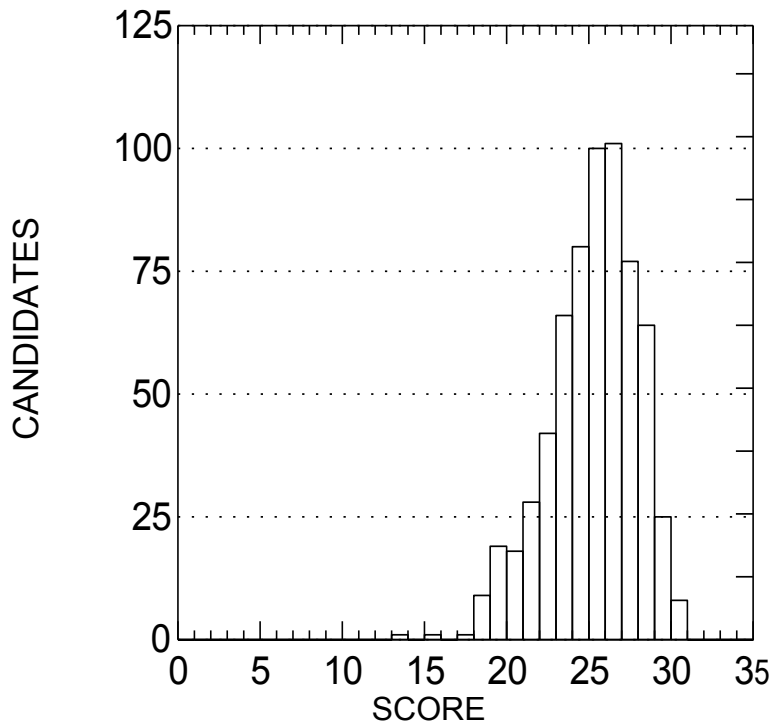
PERIODONTAL FORM 4



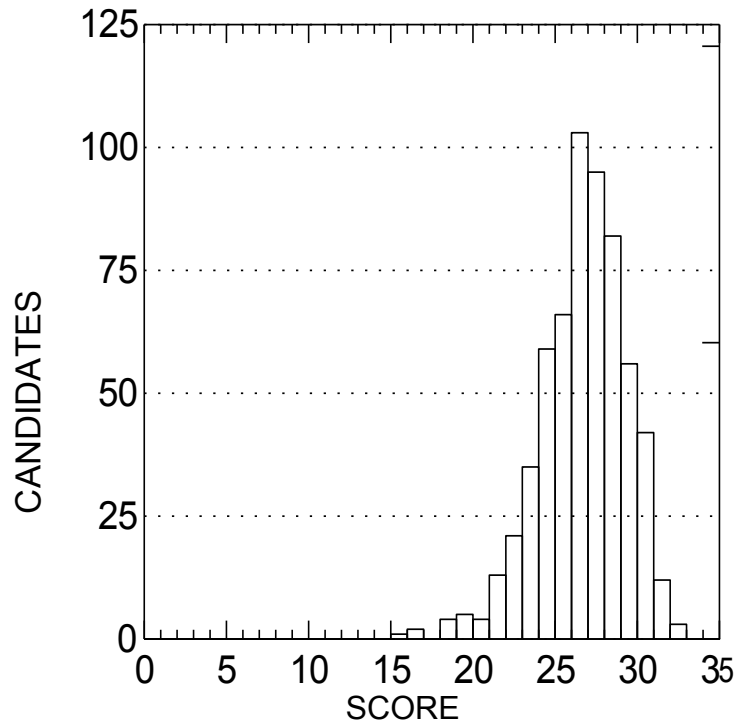
PROSTHODONTICS FORM 1



PROSTHODONTICS FORM 2



PROSTHODONTICS FORM 3



PROSTHODONTICS FORM 4

CSW has also used decision consistency estimates, which might be a better measure for supporting the validity of using the CSW tests for licensing decisions.

Decision consistency estimates are methods that evaluate the extent to which a test user can be confident in the classification decision that results from an examinee's score relative to the cut score. CSW employs the Hanson and Brennan (1990) method for estimating decision consistency. This methodology relies on information from a single administration of the test and estimates decision consistency, using variations of half-tests to estimate the level of confidence in the pass-fail decision on the full-length test. Values on decision consistency estimates range from 0.00 to 1.00; however, minimum values of 0.85 or higher are desirable. This is true particularly when there is only one pass-fail decision. In these instances it is important to be very confident in the decision that is made about candidates' pass-fail status on each test. In previous years Table 1 included the decision consistency values that met this minimum quality criterion for each form of the CSW periodontal and CSW prosthodontics tests. The 2009 reliability values were similar to the previous years' values, indicating that the decision reliability values would also be similar. Consequently, the decision consistency values were not generated for this year.

Finally, at the item (test question) level, a series of classical test theory item analyses are conducted to evaluate whether items should be retained, revised, or removed from the pool. The two diagnostic characteristics we use to review these items are the item difficulty and item discrimination. Item difficulty is the proportion of candidates who respond correctly to the item. For items to provide the greatest information to support reliability estimates, item difficulties generally range from 0.30-0.90 (i.e. 30% to 90% correct). As stated previously, items that most candidates get wrong are probably

defective items. CSW generally retains items that are above 59% correct. In a few cases, items that are more difficult are retained when the committee of subject matter experts is certain the items are correctly worded and represent important concepts. Items that most candidates get correct do not provide much information to support reliability estimates, but are generally retained by the committees because of the importance of the content.

The second item diagnostic, item discrimination, is an estimate of how high scoring candidates perform on an item relative to low scoring candidates. For an item to be useful to the pass-fail decision, it should help us distinguish between candidates who score high on the exam and those who score low. CSW uses a point-biserial correlation that is calculated between the item performance and the total score. Values on the item discrimination analyses range from 0.00 to 1.00 with a value of 1.00 being the most discriminating. The CSW test items that are important to include in the tests, but are easy for many candidates, generally have low discrimination values. These items are not removed from the tests if they have a low, but positive discrimination value. Higher difficulty items (easier items) generally are less discriminating, but are also retained if the discrimination value is positive. These items provide a negative effect on the reliability estimations and lower the average discrimination values for the tests. This is not problematic for CSW since these effects are anticipated and because of the decision to use important items without regard to the effect on the (post-test) test performance computations. Thus, the average item discrimination values shown in Table 1 are considered satisfactory.

References

- Hanson, B. F. & Brennan R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, 27(4), 345-360.
- Livingston, S. & Zieky, M. (1982), *Passing Scores*, Princeton, NJ: Educational Testing Service.
- Kolen, M. J. & Brennan, R. L. (1995). *Test Equating: Methods and Practices* (pp.35-47), New York, NY: Springer-Verlag.
- Livingston, S., (2004), *Equating Test Scores*, Princeton, NJ: Educational Testing Service.

Table 1. Statistics for the Periodontal and Prosthodontic Exams for 2009 Testing

	Perio 1	Perio 2	Perio 3	Perio 4	Pros 1	Pros 2	Pros 3	Pros 4
Descriptive statistics								
Number of items	42	39	44	46	38	28	31	33
Number of examinees	624	642	640	603	624	642	640	603
Cut score	28	27	32	34	25	18	21	22
Scale mean	36.33	33.23	38.45	40.52	30.57	22.94	25.75	27.19
Scale SD	3.39	3.42	3.46	3.30	3.22	2.75	2.68	2.65
Mean Biserial	0.28	0.32	0.27	0.24	0.19	0.27	0.18	0.17
Reliability								
KR-20 (alpha)	0.64	0.65	0.64	0.60	0.49	0.53	0.47	0.43
SEM	2.05	2.03	2.08	2.09	2.31	1.9	1.95	2.00