BOC

BOARD OF CERTIFICATION

Be Certain.™

Examination Review for 2010-2011 Testing Year

Board of Certification (BOC) Certification Examination for Athletic Trainers

Stephen B. Johnson, Ph.D.
Castle Worldwide, Inc.
Prepared March 2010

Castle™

Sound Solutions *Responsive Service*

TABLE OF CONTENTS

FIGURES

TABLES

INTRODUCTION

The Board of Certification (BOC) is a nonprofit credentialing agency that provides certification for the athletic training profession. The BOC was incorporated in 1989 to govern the certification program, which had then existed for nearly 20 years, for entry-level athletic trainers and recertification standards for certified athletic trainers. The entry-level certification program is designed to establish a common benchmark for entry into the athletic training profession. The BOC serves the public interest by developing, administering, and continually reviewing a certification process that reflects current standards of practice in athletic training.

In order to develop a credible and valid examination, the BOC contracts with Castle Worldwide, Inc. (Castle), for the design, development, and delivery the BOC's athletic trainer certification examinations. Castle follows and recommends widely accepted standards and regulations (e.g., *Standards for Educational and Psychological Testing*, American Educational Research Association, 1999; *Uniform Guidelines on Employee Selection Procedures*, EEOC, 1978; *Standards for the Accreditation of Certification Programs*, National Commission for Certifying Agencies, 2005) for the development and analysis of the BOC's athletic trainer certification examinations.

The major objective of the BOC's athletic trainer certification program is to establish that individuals have the knowledge and skills necessary to create and provide safe and effective athletic training services. It provides assurance that a certified athletic trainer has met eligibility criteria addressing training, experience, and the knowledge and skills necessary for competent performance of his or her work.

In order to attain certification, an individual must complete an entry-level athletic training education program accredited by the Commission on Accreditation of Athletic Training Education (CAATE) and pass the BOC certification examination. In order to qualify as a candidate for the BOC certification examination, an individual must meet the following requirements:

- Endorsement of the examination application by the recognized program director (PD) of the CAATE accredited education program.
- Proof of current certification in emergency cardiac care (ECC)
  (Note: ECC certification must be current at the time of initial application and any subsequent exam retake registration.)

Description of the Examination

The BOC certification examination is designed to test an individual's knowledge across the practice of athletic training based on a defined test blueprint. The examination is based on test content specifications established in a 2004 role delineation study [RD5]. From the study, six performance domains (i.e., major areas of responsibilities or duties) were established:

1. Prevention
2. Clinical Evaluation and Diagnosis
3. Immediate Care
4. Treatment, Rehabilitation, and Reconditioning;
5. Organization and Administration; and
6. Professional Responsibility

This required the development of test items and examination forms to meet these specifications and subsequently required performance standards for the examination. The examination blueprint is contained in Appendix A.

Format of Items on the Examinations

Items on the BOC certification examination consist of multiple-choice, multi-select, and drag-and-drop items.

The multiple-choice item format contains a stem and five possible response options. The stem is typically a direct question. Of the response options, there is one correct or clearly best answer, referred to as the key. The incorrect response options are called distractors.

The multi-select items contain a stem and an *N* number of possible response options. Of these options, more than one can be correct. Candidates can select more than one option. Points for an item are provided for correctly selecting an option.

The drag-and-drop items contain a stem, a list of *N* options, and a series of *N* "buckets." Candidates can select from one to *N* options from the list and "drop it" in to one of the "buckets." Options can be used once or used multiple times depending on the item. Candidates are informed of the options usability. This item is scored one point for each correctly selected option. All items are converted into a scale of 0 to 1.

Items on the BOC certification examination also are provided in a focused testlet (FT) format. These testlets are designed to assess more complex decision-making skills required for the role of an entry-level certified athletic trainer through cases that are rich in relevant, realistic, and specific information. Each FT consists of a scenario followed by five questions utilizing the range of item types. The testlets focus on questions that ask candidates to:

- Identify important facts specifically stated in the material;
- Understand the meaning of key words and phrases in the material;
- Draw conclusions and infer meanings from the material;
- Consider and evaluate evidence to support or reject different ideas; and/or
- Apply information presented in the material to a new or different situation.

This testlet format is used by many organizations and is best known for its use in reading comprehension examinations (e.g., LSAT® and GMAT®). The concept of a focused testlet is exemplified by the Medical College Admission Test (www.aamc.org/students/mcat/) and the Royal Australian College of General Practitioners (RACGP http://www.racgp.org.au/exam).

Items are constructed using guidelines established by the BOC for the development and review of items.

Delivery of the Examination

The 2010 BOC certification examination test forms included 125-scored items and 50 experimental items. Each test form included four focused testlets.

Examinations are completed in one session and candidates are allotted a period of four hours. Short tutorials are available prior to the start and a short satisfaction survey appears following the end of the examination. The BOC uses Castle's Internet-based test delivery system (PASS) for test administration.

For the 2010-2011 testing year, the examination was provided in five 14-day test windows: March/April, May/June, July/August, November, and January/February. The BOC certification examination forms consist of scored and experimental items with scored items in common with an anchor form. Candidates who fail are not restricted in their retakes during the testing year.

Number of Test Forms

One set of 125-scored items was assigned five different experimental sets for the year, creating six different test forms. Two test forms (3630 and 3631) were administered in April 2010, two (3632 and 3633) in June 2010, one in August (3634), and the final form (3635) was administered to candidates in both November 2010 and February 2011.

Equating Test Forms

Upon introduction of a new test form, the performance of candidates on the new form is equated to performance of candidates on a prior test form. The BOC equating follows the protocols for common items non-equivalent groups design using the Levine True Score Method Applied to Observed Scores with internal anchors (Kolen & Brennan, 2004). This design compares the performance of one group of test takers on one examination form to another group of test takers on an earlier examination form with a known cut-score. Ultimately, all equating is compared to the performance standard established for the base form (342) used for the current role delineation/practice analysis.

Since the original performance standard was established in 2005 on Form 342, the protocol for equating was to equate the current test forms to a form used within the last two years in order to avoid item overexposure through repeated selection of the standard setting examination versions, the removal of outdated or inappropriate items, and a potential shift over time of candidate demographics and experiences that impact the performance.

Upon administration of Forms 3630 and 3631, each of the two forms was separately equated to Form 3622, which was administered in June 2009 to assure similar results were obtained. This form was chosen as the candidates were most similar to the April 2010 administration, consisting of a large number of first-time test takers, was recently administered, and had a large pool of potential new items for equating purposes. Data for 841 candidates for Form 3622 was compared to the performance data for 957 candidates for Form 3630 and 990 for Form 3631. A total of 67 items (54%) were common between Forms 3622 and 3630/3631.

Use of Scaled Scores

Since examination forms are possibly of different difficulty, providing raw scores can be misleading. As a result, many programs, including the ACT® and SAT® examinations, use scaled scores. Scaled scores are particularly useful at providing the basis for long-term, meaningful comparisons of results across different administrations of an examination.

Scaled scores are used because, over the life of every testing program, there are situations when changes in test length occur: a decision is made to assess more or fewer areas, the numbers of items that are scored versus unscored (experimental) changes, or different examination forms of different difficulty are being compared.

For scaled scores, the passing standard (number correct) on any examination form is always reported as the same scaled score.

The equated scores for the BOC certification examination are converted via linear transformation so that the passing standard for all test forms are reported to candidates as 500 on a scale of 200 to 800.

Score Reporting

The BOC provides scaled scores and pass/fail decisions to candidates approximately two weeks after closure of a test window. Candidates pass or fail based on their scaled score performance compared to a criterion-referenced performance standard.

Examination Development

During the 2005-2006 testing year, new test specifications and the associated passing standard were introduced. All later forms of the BOC certification examination are equated back to this standard. During the 2005-2006 and 2006-2007 testing years, the examinations consisted of three components (simulation, multiple-choice, and practical), of which candidates were required to pass all three components. In 2006, the BOC undertook a major initiative to computerize its examinations and combine the practical and simulation examinations of the certification with the multiple-choice test. The three-part BOC certification examination was phased out at the beginning of the 2007-2008 testing year, and a combined two-part test was implemented in 2007.

During the 2008-2009 testing year, a decision was made to integrate the use of scenarios and alternative item types into one presentation driver and discontinue the development and use of the hybrid problems. The hybrid component was replaced by focused testlets for the April 2009 test administration.

Three examination development meetings were held in 2010 (February, July, and November), with the groups focused on building a new examination forms, reclassifying the item bank to the new role delineation content outline [RD6], reviewing and developing stand-alone items, and reviewing and developing focused testlets.

Stand-Alone Committee

During the 2010 meetings, the committee reviewed 150 experimental items administered during March/April, May/June, and July/August, modifying and approving as necessary for further testing as experimental or scored items. The committee also reviewed and developed 138 stand-alone items for field-testing and updated references for items.

Focused Testlet Committee

The committee reviewed 12 experimental focused testlets; approved 14 focused testlets for field-testing; and developed, reviewed, and/or updated 14 more focused testlets.

Test Form Assembly

As part of RD6, a new test blueprint was developed and approved. During the November 2010 meeting, two sets of scored items [Scored Set 1 and Scored Set 2 aligned to the new test blueprint] were assembled for administration during the 2010-2011 testing year. Scored Set 1 was reviewed in January 2011 in preparation for the passing standard meeting in February 2011. Following the April 2011 test administration, a set of common/equating items from Scored Set 1 will be identified and combined with Scored Set 2.

I-Dev

The BOC also uses Castle's I-Dev system for online development of multiple-choice items. In 2010, BOC subject matter experts completed development of 513 validated items.

Item Bank

Currently, BOC has 650 items [multiple-choice, multi-select, and drag-and-drop] in its certification examination item bank. In addition, 358 items [multiple-choice, multi-select, and drag-and-drop] are included in the self-assessment item bank and 27 items [multiple-choice, multi-select, and drag-and-drop] are included in the sample test item bank. Finally, 125 items [multiple-choice] are included in the recertification assessment item bank.

Castle staff continually reviews and edits items and the resulting examination forms for psychometric and publication purposes. Items for the examination are stored in I-Bank, Castle's proprietary item-banking system.

ANALYSIS OF THE EXAMINATION

Candidate Performance

Statistics reported refer to the performance of 'analyzed' candidates for the BOC certification examination. Statistical reports are generated for a particular time (e.g., a testing window). Some candidates are excluded from the pool of analyzed data, specifically those candidates who completed less than 25% of their examinations. It is likely that these candidates experienced problems, such as being late to the site or other issues, and therefore, their data is problematic. As of 2007, the three cohorts of candidates reported for the BOC examinations are:

1. First-time candidates – candidates reported as first-time test takers and/or recent college graduates from athletic training education programs accredited by the CAATE.
2. Retakes – candidates who re-sat the examination one or more times.
3. All – candidates who tested.

Candidates Excluded from this Report

The report does not include, except where noted, those candidates who were administered the BOC certification examination via paper-and-pencil or those candidates with incomplete data. As a result, the number of candidates analyzed for this report may not match the number of candidates who sat for the BOC athletic trainer examination. Data from previous years may only include two of the three cohorts.

Data for individual tables also may differ due to exclusion of some candidates from the analysis for that table. Data prior to the introduction of the two-part examination (April 2007) are excluded from the remainder of this report, except where noted, because the program used to assess candidates is not equivalent to the revised BOC certification examination.

There were 5,711 reported administrations of the BOC examination during the 2010-2011 testing year, a drop of approximately 8% compared to both 2009-2010 (6,171) and 2008-2009 (6,135) testing years. Of the 5,711 results, 2,963 (52%) were administered to first-time candidates, an increase from 2008-2009 and 2009-2010 (46%).

Pass Rates

From the 2007-2008 testing year onward, candidates were required to pass BOC certification examination as documented above. Table 1 provides annual pass rates for BOC administrations from 2005-2006, but only reports the pass rates for 2005-2006 and 2006-2007 that are associated with the multiple-choice element.

Table 1: Number of Candidates in Three Cohorts and Pass Rate for BOC Examinations, 2005-2006 to 2010-2011 (2005-2006 and 2006-2007 are for the Multiple-Choice Element Only).

Year	First-time	Pass	% Pass	Retake	Pass	% Pass	All	Pass	% Pass
2005-2006	2,074	968	46.7%	3,017	660	21.9%	5,091	1,628	32.0%
2006-2007	2,322	1,125	48.4%	3,549	1,076	30.3%	5,871	2,201	37.5%
2007-2008	1,495	584	39.1%	3,196	1,073	33.6%	4,691	1,657	35.3%
2008-2009	2,762	1,423	51.5%	3,373	1,035	30.7%	6,135	2,458	40.1%
2009-2010	2,852	1,235	43.3%	3,319	1,120	33.7%	6,171	2,355	38.2%
2010-2011	2,963	1,800	60.7%	2,748	938	34.1%	5,711	2,738	47.9%

The three-component BOC certification examination resulted in a pass rate for first-item test takers of 26.2% in 2005-2006 and 31.5% in 2006-2007. This was substantially lower than the pass rates for the combined examination protocol used since 2007-2008.

A test of proportions indicated that the pass rate for all examinations administered in 2010-2011 is higher than the percentage that passed the 2009-2010 examination (z = 5.10, p < .05) and significantly higher than 2008-09 (z = 8.55, p < .05). The pass rate for retake candidates was not significantly different from 2009-2010 (z = 0.33, p > .05). The pass rate for first-time candidates was higher for 2010-2011 compared to 2009-2010 (z = 13.28, p < .05) and 2008-09 (z = 7.01, p < .05).

Table 2 details the pass rates for each form by testing window.

Table 2: Passing Rates for Each Test Form for All Candidates for BOC Examinations, 2010-2011.

	Frequency			Percent	
Form	Fail	Pass	Total	Fail	Pass
3630	403	554	957	42.1%	57.9%
3631	407	582	989	41.2%	58.8%
April	*810*	*1136*	*1946*	*41.6%*	*58.4%*
3632	379	367	746	50.8%	49.2%
3633	404	348	752	53.7%	46.3%
June	*783*	*715*	*1498*	*52.3%*	*47.7%*
3634	538	349	887	60.7%	39.3%
August	*538*	*349*	*887*	*60.7%*	*39.3%*
November	475	276	751	63.2%	36.8%
February	363	261	624	58.2%	41.8%
3635 Total	*838*	*537*	*1375*	*60.9%*	*39.1%*
ALL	**2969**	**2737**	**5706**	**52.0%**	**48.0%**

Distribution of Candidate Scores

Table 3 details the overall scaled score performance for the BOC certification examination for 2010-2011 with a comparison of the performance of 2008-2009 candidates.

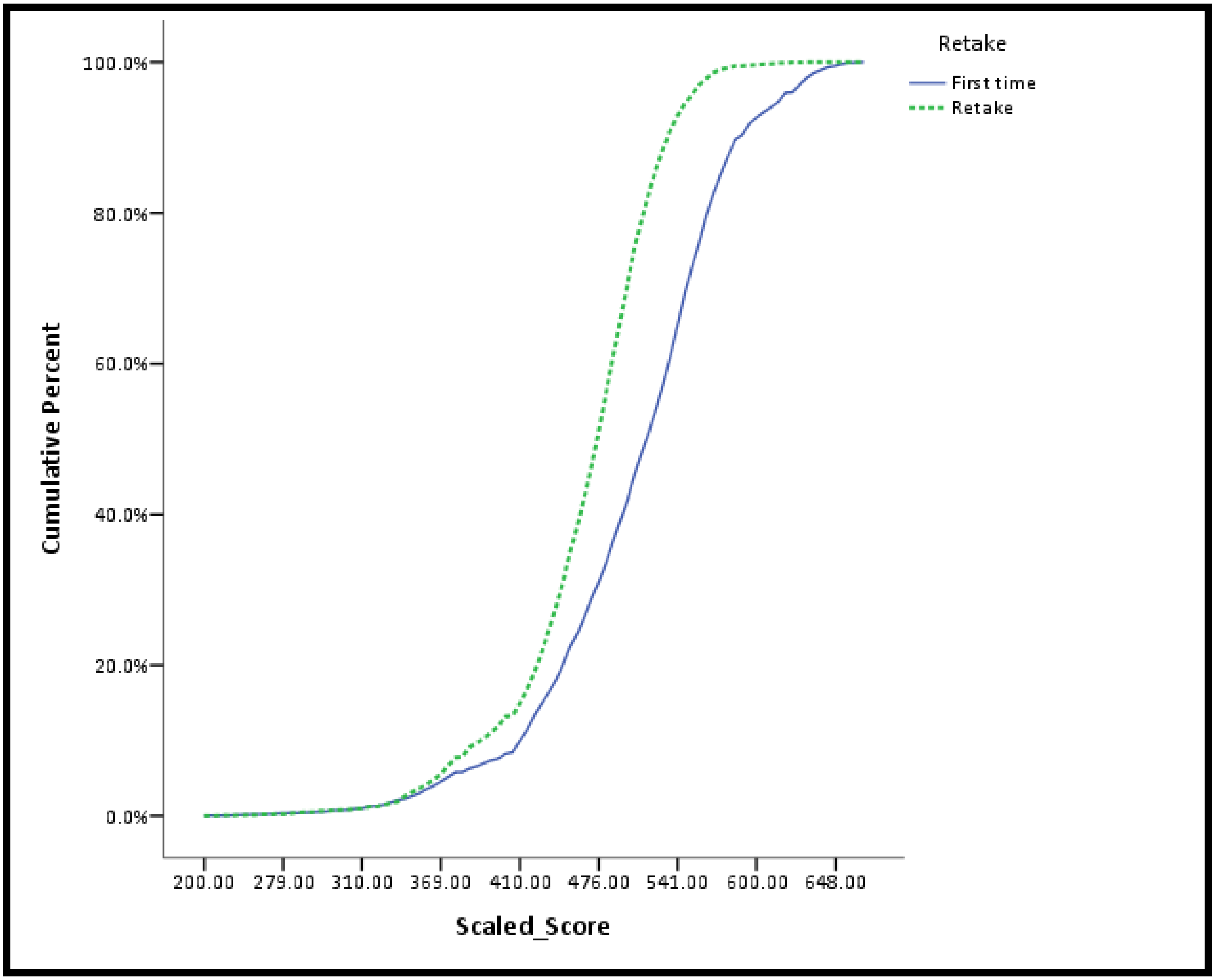
Table 3: Number of Candidates in Three Cohorts, Minimum, Maximum and Average Scaled Score, Median and Mode Scaled Score, and Standard Deviation (Scaled Score) for BOC Examinations, 2010-2011.

Cohort	N	Avg.	Median	Std. Dev.	Min	Max
First-time	2,963	508	517	71	200	672
Retake	2,748	470	476	56	220	624
All 2010-2011	5,711	490	494	67	200	672
All 2009-2010	6,171	476	482	58	200	638
All 2008-2009	6,135	473	476	79	200	686

A Univariate General Linear Model (GLM) test determined that there was a statistically significant difference in the scaled scores of retake and first-time candidates [F (1, 5709) = 493.6, p < .001, η = .08]. The difference was much greater than determined from the previous two testing years. In 2009-2010 there was a 10 scale point difference in the average between the two groups. This increased to a 41 point difference in 2010-2011.

For 2010-2011, the score distributions of first-time and retake candidates were almost identical despite this small statistical difference. This was confirmed by a review of the distribution of scaled scores for first-time and retake candidates for whom no difference in the score distributions can be noted (Figure 1).

Figure 1: Cumulative Percentage of First-time New Graduates and Retake Candidates by Scaled Score, BOC 2010-2011.



Test Form Summary Statistics

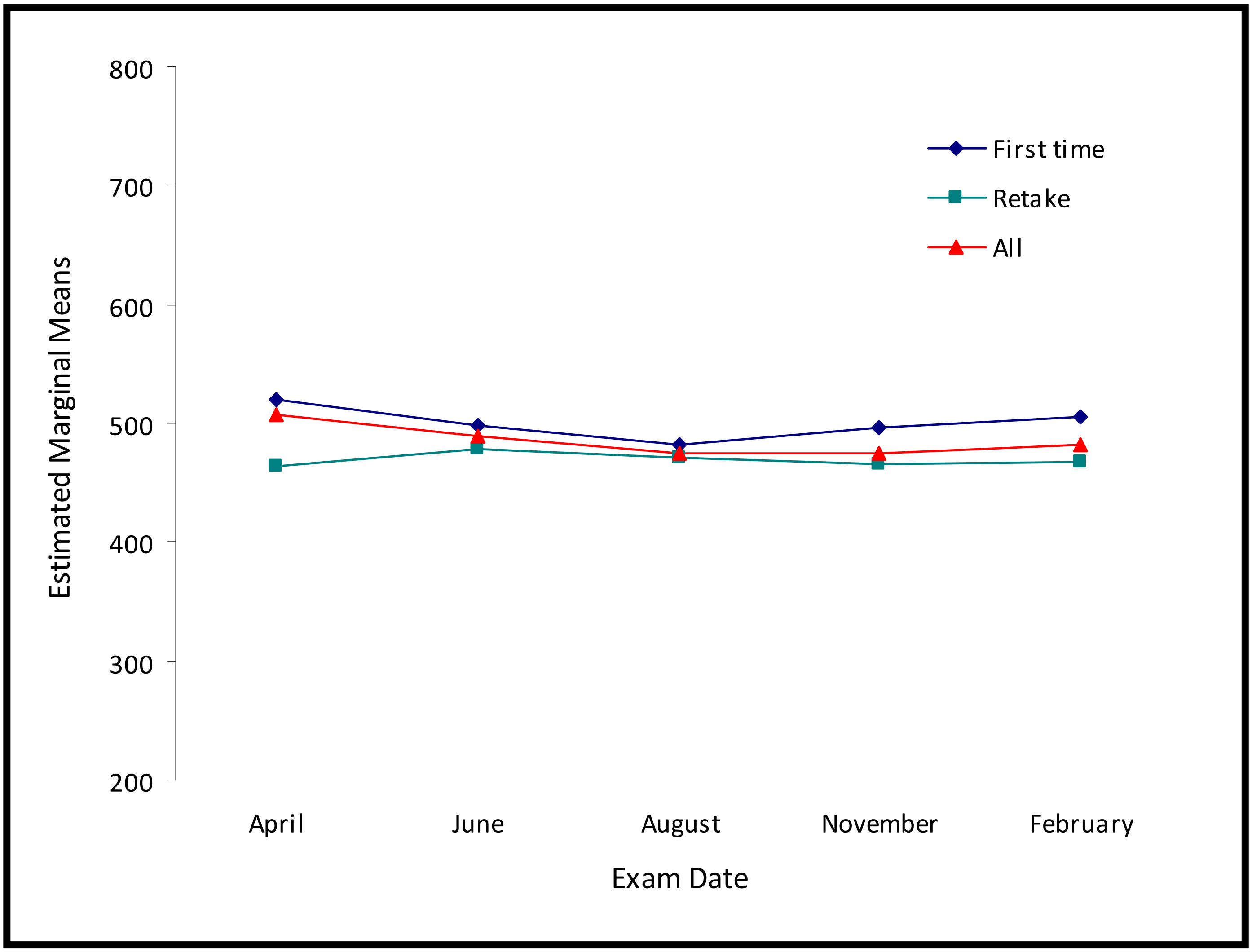
Table 4 provides form descriptive statistics for each testing window (see Appendix A for information on the statistics reported).

Table 4: Summary Test Form Statistics in Scaled Scores for All Candidates for BOC Examinations, 2010-2011.

Exam	N	Mean	Median	Std. Dev.	Min	Max
3630	957	508	511	68	220	672
3631	989	505	511	70	215	666
April	*1946*	*506*	*511*	*69*	*215*	*672*
3632	746	491	494	66	256	642
3633	752	487	494	66	244	648
June	*1498*	*489*	*494*	*66*	*244*	*648*
3634	890	474	482	62	244	654
August	*890*	*474*	*482*	*62*	*244*	*654*
November	751	475	482	64	244	660
February	626	482	488	67	200	654
3635 Total	*1377*	*478*	*482*	*65*	*200*	*660*
ALL	**5711**	**490**	**494**	**67**	**200**	**672**

As shown in Table 4, there is some difference in the scaled scores for each testing window and for each form of the BOC certification examination. A statistical test (a Univariate General Linear Model) was conducted to examine whether there was any statistical difference in the scaled scores for candidates based on the month they tested and whether they were retake or first-time candidates. There was a significant, though small, interaction between the candidate's month and retake status [F(4,5710) = 20.88, p =<.05, η = .01]. Figure 2 provides a comparison of scaled score means for the groups over the testing year.

Figure 2: Scaled Score Means for First-time, Retake, and All Candidates for Each of Five Testing Windows, BOC 2010-2011.



Difficulty and Discrimination

Average difficulty and discrimination was computed for all test forms. Table 5 contains the average, minimum, and maximum values for difficulty and discrimination. Additional summary data are contained in Appendix B.

Table 5: Average Difficulty and Discrimination Statistics for Scored Items by Test Form for All Candidates for BOC Examinations, 2010-2011.

Test Form	Statistic	Candidates	Average	Min	Max
3630	Difficulty	957	0.71	0.14	1.00
April	Discrimination	957	0.17	-0.02	0.44
3631	Difficulty	989	0.71	0.14	1.00
April	Discrimination	989	0.18	0.02	0.44
3632	Difficulty	746	0.69	0.12	1.00
June	Discrimination	746	0.17	-0.03	0.39
3633	Difficulty	752	0.69	0.12	0.99
June	Discrimination	752	0.17	-0.02	0.42
3634	Difficulty	887	0.67	0.11	1.00
August	Discrimination	887	0.15	-0.03	0.35
3635	Difficulty	751	0.67	0.12	1.00
November	Discrimination	751	0.15	-0.04	0.41
3635	Difficulty	624	0.68	0.13	1.00
February	Discrimination	624	0.16	-0.06	0.41
ALL	**Difficulty**	5706	0.69	0.11	1.00
	Discrimination	5706	0.17	-0.06	0.44

Overall, discrimination statistics for the items were within an acceptable range of 0.1 to 0.3. Average difficulty for the BOC certification examination forms was appropriate. Data on the range of difficulty and discrimination statistics obtained for first administration of the 125-scored items is contained in Table 6 below.

Table 6: Summary of Item Discrimination and Difficulty for First Forms Tested, BOC 2010-2011.

Form		Difficulty		Discrimination	
3630	**Average**	**0.71**		**0.17**	
	Median	**0.73**		**0.22**	
	Min	**0.14**		**-0.02**	
	Max	**1.00**		**0.44**	
		Frequency	*Percent*	*Frequency*	*Percent*
	< 0.0			1	0.8
	0.0 to 0.1			13	10.4
	0.1 to 0.2	2	1.6	37	29.6
	0.2 to 0.3			49	39.2
	0.3 to 0.4	3	2.4	23	18.4
	0.4 to 0.5	6	4.8	2	1.6
	0.5 to 0.6	20	16		
	0.6 to 0.7	25	20		
	0.7 to 0.8	21	16.8		
	0.8 to 0.9	32	25.6		
	> 0.9	16	12.8		
	Total	125	100	125	100
3631	**Average**	**0.71**		**0.18**	
	Median	**0.73**		**0.23**	
	Min	**0.14**		**0.02**	
	Max	**1.00**		**0.44**	
		Frequency	*Percent*	*Frequency*	*Percent*
	< 0.0				
	0.0 to 0.1			10	8.0
	0.1 to 0.2	1	0.8	36	28.8
	0.2 to 0.3	1	0.8	55	44.0
	0.3 to 0.4	4	3.2	22	17.6
	0.4 to 0.5	4	3.2	2	1.6
	0.5 to 0.6	21	16.8		
	0.6 to 0.7	26	20.8		
	0.7 to 0.8	24	19.2		
	0.8 to 0.9	28	22.4		
	> 0.9	16	12.8		
	Total	125	100	125	100

Domain Performance

Test validity is a concept that refers to how well an examination measures what it is designed to measure. Test forms for the BOC certification examination were constructed according to test specifications that were based on the results of the 2004 role delineation study [RD5]. This study was undertaken to define the job-related activities, knowledge, and skills required of entry-level athletic trainers. To ensure that test items account for the content areas presented in the test specifications, each item has been classified by content experts according to its application to the practice domains and tasks of the role delineation study.

Each test item has been linked to a specific content area of the test specifications, and items meet minimum standards of criticality to entry-level work as an athletic trainer. Thus, the procedures used to construct the test support the inference that the test has been built to achieve its stated purpose. Consistent with the objectives of the BOC examination program, the test is designed to separate candidates into two distinct groups: candidates whose knowledge and skill levels are deemed acceptable for entry-level certification as a practitioner and candidates whose level of knowledge falls below the minimum requirements for certification. Test forms for the BOC certification examination are not intended as predictors of future success within the profession.

There are six performance domains in the content framework for the BOC examination, consistent with the role delineation study upon which the examination is based. The domains are *Prevention; Clinical Evaluation and Diagnosis; Immediate Care; Treatment, Rehabilitation, and Reconditioning; Organization and Administration; and Professional Responsibility.* Table 7 reports descriptive statistics at the domain level using raw score.

Table 7: Domain Level Statistics for Each Test Form for All Candidates for BOC Examinations, 2010-2011 (Based on Raw Scores).

Form	Statistic	Prevention	Clinical Evaluation and Diagnosis	Immediate Care	Treatment Rehabilitation and Reconditioning	Organization and Administration	Professional Responsibility
3630	N	957					
April	Minimum	6	7	2	7	4	3
	Maximum	20	28	22	29	14	11
	Mean	13.9	20.8	15.5	20.5	10.5	8.2
	Std. Deviation	2.4	3.5	2.8	3.5	1.8	1.5
3631	N	989					
April	Minimum	4	8	5	7	2	2
	Maximum	20	29	22	29	14	11
	Mean	13.9	20.5	15.4	20.4	10.6	8.2
	Std. Deviation	2.5	3.5	2.8	3.8	1.9	1.5
3632	N	746					
June	Minimum	4	8	4	8	4	3
	Maximum	19	28	21	29	14	11
	Mean	13.3	20.1	15.0	19.8	10.2	8.0
	Std. Deviation	2.4	3.7	2.7	3.6	1.8	1.6
3633	N	752					
June	Minimum	6	6	7	8	3	3
	Maximum	19	28	22	29	14	11
	Mean	13.2	19.9	14.8	19.6	10.2	8.1
	Std. Deviation	2.3	3.5	2.7	3.5	1.9	1.6
3634	N	887					
August	Minimum	7	9	6	7	4	3
	Maximum	19	27	21	28	14	11
	Mean	12.9	19.3	14.5	19.0	10.3	7.6
	Std. Deviation	2.3	3.4	2.7	3.5	1.8	1.5
3635	N	751					
November	Minimum	7	8	5	8	4	2
	Maximum	20	27	21	28	14	11
	Mean	13.0	19.6	14.4	19.1	10.1	7.7
	Std. Deviation	2.4	3.4	2.8	3.5	1.8	1.5
3635	N	624					
February	Minimum	6	6	3	4	7	3
	Maximum	20	28	22	27	14	11
	Mean	13.2	19.9	14.6	19.6	10.1	7.7
	Std. Deviation	2.4	3.4	2.9	3.6	1.8	1.6

Correlations in candidate performance between the six domains ranged from 0.24 to 0.56, indicating that the domains were assessing somewhat different constructs (see Appendix B). These correlations are consistent with the results obtained for 2008-2009 and 2009-2010.

Test Form Internal Reliabilities

Reliability is assessed using the Brennan-Kane statistic (Brennan & Kane, 1977), a measure typically used for estimating decision consistency for criterion referenced tests, and the Standard Error of Measurement (presented in Scaled Score units). The Brennan-Kane reliability estimate accounts for the more constrained dispersion of candidate scores and the use of a passing standard and is consistent with reporting standards for accreditation purposes (Table 7).

Table 8: Internal Reliability Estimates for Each Test Form for All Candidates for BOC Examinations, 2010-2011.

Form	N	Std. Error	Brennan-Kane Estimate
3630	957	16.54	0.94
3631	989	19.74	0.92
3632	746	13.27	0.96
3633	752	13.14	0.96
3634	887	13.82	0.95
3635	751	14.34	0.95
3635	624	14.89	0.95
ALL	**5706**	**15.50**	**0.95**

Data presented in Table 8 show that each testing window meets general guidelines for a Brennan-Kane statistics of greater than 0.70 and is consistent with previous years. Standard Errors of Measurement also are consistent with previous years.

Summary Test Form Data

Data presented in the following table summarize the performance of the test forms used for the BOC certification examination and are consistent with reporting requirements for NCCA/ICE Accreditation (Table 8). The data also is presented for each form that represents the common set of 125 scored items.

Table 9: Summary Statistics for the 2010-2011 Administrations of BOC Athletic Trainer Test Forms.

Form #	Total # of Candidates Tested	% of Candidates Passing Each Form	Passing Point	Average Score	Standard Deviation	Standard Error of Measurement	Reliability Estimate	Total # of Items on Form
3630	957	58%	500	508	68	16.54	0.94	175
3631	989	59%	501	505	70	19.74	0.92	175
3632	746	49%	502	491	66	13.27	0.96	175
3633	752	46%	503	487	66	13.14	0.96	175
3634	887	39%	504	474	62	13.82	0.95	175
3635	751	37%	505	475	64	14.34	0.95	175
Total	**5706**	**48%**		**490**	**66**	**15.50**	**0.95**	

Data presented Table 9 is in scale score units for passing point, average score, standard deviation, and standard error of measurement.

CONCLUSION

Statistics concerning the quality of the BOC certification examinations as a measurement device indicate that the examination complies with psychometric requirements that pertain to certification and licensure tests. Notably, estimates of reliability and equivalence across forms for the various parts of the examination are strong. Likewise, candidate performance on all parts of the examination is consistent with the public protection mission of the BOC.

REFERENCES

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

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277–289.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16,* 297–334.

Equal Employment Opportunity Commission (EEOC), U.S. Civil Service Commission, U.S. Department of Labor, and U.S. Department of Justice. (1978). Uniform Guidelines on Employee Selection Procedures. *Federal Register, 43 (166)*, 38290-38315.

Kolen, M.J., & Brennan, R.L. (2004) Test *Equating, Scaling and Linking: Methods and Practices Statistics for Social Science and Behavioral Sciences* (2 ed.). Springer-Verlag New York Inc.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

APPENDICES

Appendix A: Definitions of Form Statistics

Mean Score

Average score of the analyzed candidates. It is the sum of all the analyzed candidate scores divided by the total number of analyzed candidates.

Standard Deviation

The standard deviation describes the amount of spread among the scores of the analyzed candidates. The larger the standard deviation, the more spread out the scores. A large standard deviation indicates that candidate scores are far from the mean, and a small standard deviation indicates that they are clustered closely around the mean. Larger standard deviations make it easier to discriminate among candidates at different score levels.

Mean scores and standard deviations are related to each other. Chebyshev's inequality shows that that for most distributions at least $(1 - 1/k^2) \times 100\%$ of the values are within k standard deviations from the mean score:

- At least 50% of the values are within 1.4 standard deviations from the mean.
- At least 75% of the values are within 2 standard deviations from the mean.
- At least 89% of the values are within 3 standard deviations from the mean.
- At least 94% of the values are within 4 standard deviations from the mean.
- At least 96% of the values are within 5 standard deviations from the mean.
- At least 97% of the values are within 6 standard deviations from the mean.
- At least 98% of the values are within 7 standard deviations from the mean.

Standard Error of Measurement

The Standard Error of Measurement (SE*m*) is used to determine the range of certainty around a candidate's reported score. The SE*m* makes it possible to determine how reliable a particular test is and how much confidence we can place in the scores it yields.

The SE*m* estimates the range of scores candidates might get if they were to take the same test over and over again (assuming no benefit from the repeated practice). The error range represents limits around an observed test score within which we would expect to find the true score. The SE*m* is used to create upper and lower boundaries around an observed score. The lower the SE*m* the more reliable to observed score is.

Min and Max (Low and High Score)

Lowest and highest score for candidates analyzed.

Avg. Diff

This refers to average item difficulty. Difficulty is an assessment of the proportion of candidates who answered items correctly; for this reason, it is frequently called the *p-value*. Difficulty ranges between 0.0 and 1.0, with a higher value indicating that a greater proportion of candidates responded to an item correctly, identifying it as an easier item. Most individual item difficulties should range from .30 (difficult) to .92 (easy).

The average item difficulty on a form is the average *p-value* across all items. The statistic can be useful in estimating how hard the test was relative to the ability level of the group. When coupled with the information about individual item difficulty (e.g., Castle's *Item Analysis Report*), this statistic can give some indication of the extent to which the test difficulty might have influenced some of the other statistical indices on the test.

For example, form reliability is typically higher when items of medium difficulty are predominant. In general, item difficulties slightly higher than medium difficulty [halfway between the probability of successfully getting an item correct by chance (e.g., .25 for a four-option item) and 1.00 (e.g., 0.63 for an examination with all four-option items)] tend to maximize both test reliability and discrimination.

Avg. Discrim.

This refers to the average item discrimination statistic for the candidates analyzed. Discrimination is a statistic that examines whether an item can discriminate between those candidates who possess the minimally acceptable level of knowledge to become certified and those candidates who do not.

There are a variety of item discrimination statistics, and Castle uses the *point-biserial correlation*. This statistic looks at the relationship between a candidate's performance on an item (correct or incorrect) and the candidate's score on the overall test. For an item that is highly discriminating, overall, the candidates who responded to the item correctly also did well on the test, whereas the candidates who responded to the item incorrectly tended to do poorly on the test. The possible range of the discrimination index is -1.0 to 1.0.

When interpreting the value of discrimination, it is important to be aware that there is a relationship between an item's difficulty and its discrimination. If an item has a very high (or very low) difficulty, the potential value of the discrimination index will be much less than if the item has a mid-range difficulty. In other words, if an item is either very easy or very hard, it is not likely to be very discriminating. Certification tests, with their often high *p-values*, may have most item discriminations in the range of 0.0 to 0.3.

Reliability Measures

Test reliability is an important statistic for any program. Reliability is the degree of consistency of a set of measurements or a measurement instrument. Reliability is typically whether the same instrument gives, or is likely to give, the same measurement (e.g., test-retest), or in the case of more subjective instruments, whether two independent assessors give similar scores (inter-rater reliability). Reliability is affected by both the number of candidates and the number of items. If the items are well constructed, the more items on a test, the more reliable the test is.

Reliability does not imply validity. A reliable measure is measuring something consistently, but the statistics does not tell us what it is measuring. As a general rule, a reliability of 0.80 or higher is desirable. The higher the reliability estimated for a test, the more confidence that a test user can have that the discriminations between candidates at different score levels on the test are stable differences.

There are numerous assessments of test reliability, Cronbach's alpha (Cronbach, 1951), Decision Consistency, Brennan Kane (Brennan & Kane, 1977), and K-R20 (Kuder & Richardson, 1937).

Appendix B: Correlations of Candidate Performance on the Six Domains

		DOMAIN_01	DOMAIN_02	DOMAIN_03	DOMAIN_04	DOMAIN_05	DOMAIN_06
Prevention	Pearson Correlation	1	.423**	.405**	.480**	.317**	.242**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	5711	5711	5711	5711	5711	5711
Clinical Evaluation & Diagnosis	Pearson Correlation	.423**	1	.520**	.557**	.386**	.244**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	5711	5711	5711	5711	5711	5711
Immediate Care	Pearson Correlation	.405**	.520**	1	.516**	.386**	.255**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	5711	5711	5711	5711	5711	5711
Treatment Rehabilitation and Reconditioning	Pearson Correlation	.480**	.557**	.516**	1	.397**	.289**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	5711	5711	5711	5711	5711	5711
Organization and Administration	Pearson Correlation	.317**	.386**	.386**	.397**	1	.276**
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	5711	5711	5711	5711	5711	5711
Professional Responsibility	Pearson Correlation	.242**	.244**	.255**	.289**	.276**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	5711	5711	5711	5711	5711	5711

**. Correlation is significant at the 0.01 level (2-tailed).