Examination Report for 2012-2013 Testing Year

Board of Certification (BOC) Certification Examination for Athletic Trainers

Stephen B. Johnson, Ph.D.
Castle Worldwide, Inc.
Prepared April 2013

Castle
Sound Solutions *Responsive Service*

TABLE OF CONTENTS

FIGURES

TABLES

INTRODUCTION

The Board of Certification, Inc. (BOC) is a non-profit 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 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 of 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 certification 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 Certification 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 certification examination is based on test content specifications established in the role delineation/practice analysis study (RD/PA6) introduced in April 2011. From the study, five performance domains (i.e., major areas of responsibilities or duties) were established:

1. Injury/illness Prevention and Wellness Protection;
2. Clinical Evaluation and Diagnosis;
3. Immediate and Emergency Care;
4. Treatment and Rehabilitation; and
5. Organization and Professional Health and Well-being.

All items and test forms are written to meet these specifications and subsequent performance standards for the certification examination.

Format of Items on the Certification Examination

The BOC certification examination consists of multiple-choice, multi-select, hotspot, and drag-and-drop items.

The multiple-choice items contain a stem and four or 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*. Points for an item are provided for correctly answering the item.

The multi-select items contain a stem and four to eight 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 hotspot items contain a stem and an image. Candidates place a "hotspot" on the correct portion of the image. Points are provided for correctly placing the hotspot.

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 into one of the "buckets." Depending on the item, options can be used once, multiple times, or not at all. 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 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 focused testlet 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 Certification Examination

The BOC certification examination test forms include 175 items (scored and experimental).

Certification 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 2012-2013 testing year, the certification examination was administered in five 14-day test windows: March/April, May/June, July/August, November, and 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

Two sets of scored items were developed for 2012-2013. Each scored set was assigned different experimental sets for the year, creating six different test forms. Forms 362(7), 362(8), and 362(11) comprised scored set A, and Forms 362(9), 362(10), and 362(12) comprised scored set B. Scored set A was first administered in April 2012 and equated to Form 362(3), which was administered in June 2011 to 1,177 candidates. Scored set A and Form 362(3) had 70 items in common. Scored set B was first administered in June 2012 and also equated to Form 362(3), sharing 65 items in common.

Standard Setting and Equating of Test Forms

In February 2011, a panel of 10 currently certified athletic trainers was convened to establish the performance standard to be implemented for the revised test blueprint (RD/PA6). The panel reviewed the scored questions for Forms 362(1) and 362(2) introduced in April 2011. The panel participated in three rounds of data collection and used a modified Angoff model, the Yes/No technique (Impara & Plake, 1997).

All later forms of the examination are equated following the protocols for common-item 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.

The protocol for equating is 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.

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 of items answered correctly) 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.

Certification Examination Development

During 2010-2011, new test specifications and the associated passing standard were introduced. All later forms of the BOC certification examination are equated back to this standard.

Since 2006, the BOC has provided a computerized certification examination. Prior to 2007-2008, the certification examination consisted of three separate components. Since that period, the certification examination has consisted of one assessment experience for candidates. During the 2008-2009 testing year, focused testlets were introduced to the testing model.

Two-day meetings for item review and test form development were held in 2012 in February, July, and November. The meetings focused on the development and review of focused testlets and the review of stand-alone items.

ANALYSIS OF THE CERTIFICATION 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 test 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 certification examinations are:

1. First-time candidates – candidates from athletic training education programs accredited by the CAATE reported as first-time test takers of the certification examination.
2. Retakes – candidates who re-sat for the certification 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 certification 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 April 2007 is excluded from the remainder of this report, except where noted, because the program used to assess candidates was not equivalent to the current BOC certification examination protocol.

There were 4,950 reported administrations of the BOC certification examination during the 2012-2013 testing year, an increase of 1% from 2011-2012 (4,886). Continuing an upward trend since 2008-2009, of the 4,950 administrations, 3,631 (73%) examinations were administered to first-time candidates, compared with 66% in 2011-12, 52% in 2010-2011 and 46% in 2008-2009 and 2009-2010.

Pass Rates

Table 1 provides annual pass rates for the BOC certification examination since 2005-2006. Data for 2005-2006 and 2006-2007 are for the multiple-choice component of the three-part assessment used by the BOC at the time. Forms prior to 2011-2012 were administered under a different blueprint and standard, and information is provided for historical purposes only.

Table 1: Number of Candidates in Three Cohorts and Pass Rates for BOC Certification Examination, 2005-2006 to 2011-2012.[1]

Year	First-time	Pass	% Pass	Retake	Pass	% Pass	All	Pass	% Pass
RD5									
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%
RD6									
2011-2012	3,222	2,653	82.3%	1,664	696	41.8%	4,886	3,269	66.9%
2012-2013	3,631	2,935	80.8%	1,319	507	38.4%	4,950	3,442	69.5%

Table 2 details the pass rates for each form by test window for the administrative year.

Table 2: Passing Rates for Each Test Form for All Candidates for BOC Certification Examination, 2012-2013.

			Frequency		Percent	
Test Window	Form	Fail	Pass	Total	Fail	Pass
April	362(7)	202	876	1,078	18.7%	81.3%
	362(8)	197	881	1,078	18.3%	81.7%
	Total	*399*	*1,757*	*2,156*	*18.5%*	*81.5%*
June	362(9)	187	350	537	34.8%	65.2%
	362(10)	201	335	536	37.5%	62.5%
	Total	*388*	*685*	*1,073*	*36.2%*	*63.8%*
August	362(11)	268	256	524	51.1%	48.9%
	Total	*268*	*256*	*524*	*51.1%*	*48.9%*
November	362(12)	259	292	551	47.0%	53.0%
	Total	*259*	*292*	*551*	*47.0%*	*53.0%*
February	362(12)	194	452	646	30.0%	70.0%
	Total	*194*	*452*	*646*	*30.0%*	*70.0%*
	ALL	**1,508**	**3,442**	**4,950**	**30.5%**	**69.5%**

[1] 2005-2006 and 2006-2007 data are for the multiple-choice component only.

Distribution of Candidate Scores

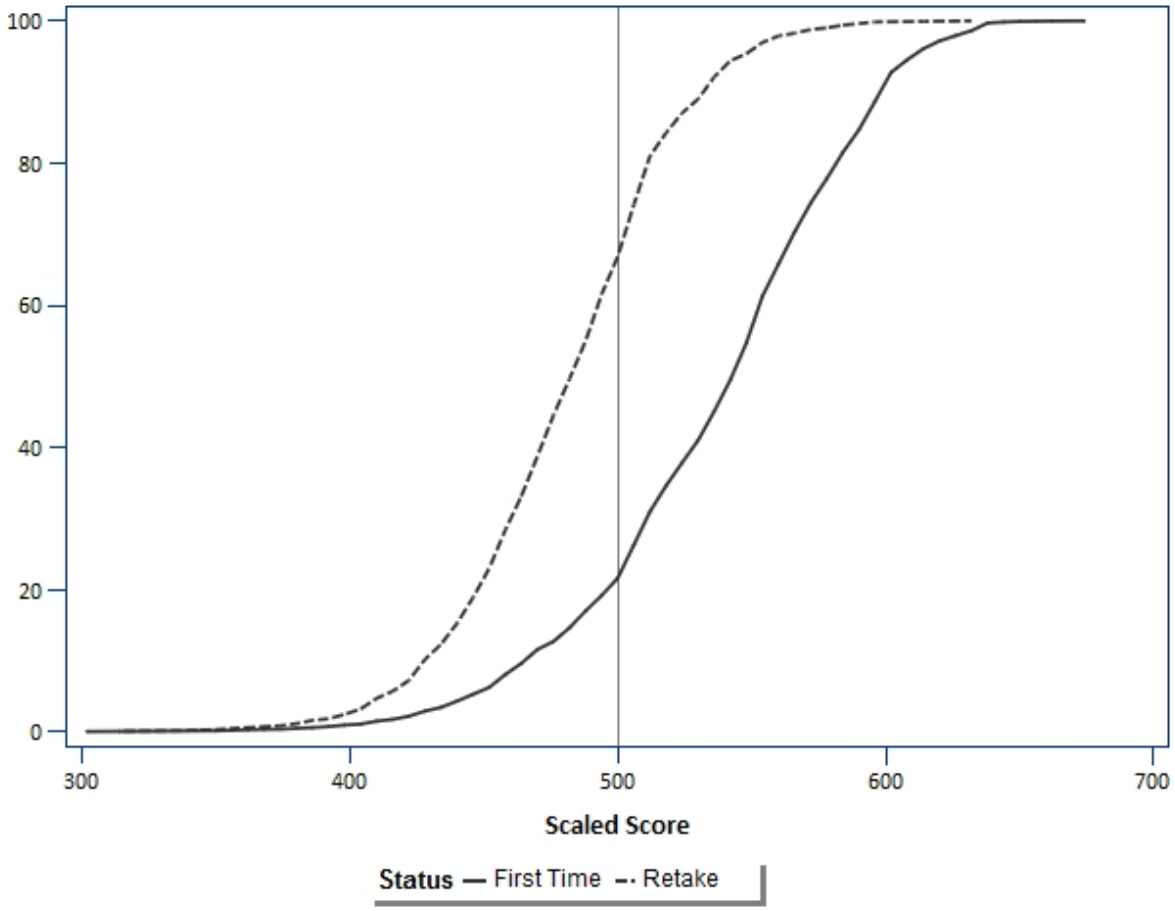
Table 3 details the overall scaled score performance for the BOC certification examination for 2012-2013 with a comparison of the performance of candidates since 2008-2009.

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 Certification Examination.

Cohort	N	Avg.	Median	Std. Dev.	Min	Max
All 2012-13	4,950	524	524	54	302	674
First-time	3,631	539	548	51	302	674
Retake	1,319	484	488	41	314	632
All 2011-12	4,886	525	524	54	230	692
First-time	3,222	542	548	51	272	692
Retake	1,664	491	494	44	230	644
All 2010-11	5,711	490	494	67	200	672
First-time	2,963	508	517	71	200	672
Retake	2,748	470	476	56	220	624
All 2009-10	6,171	476	482	58	200	638
All 2008-09	6,135	473	476	79	200	686

Figure 1 presents a cumulative frequency distribution for 2012-2013 retake and first-time candidates. The figure represents the proportion of candidates who scored at a scale score or lower.

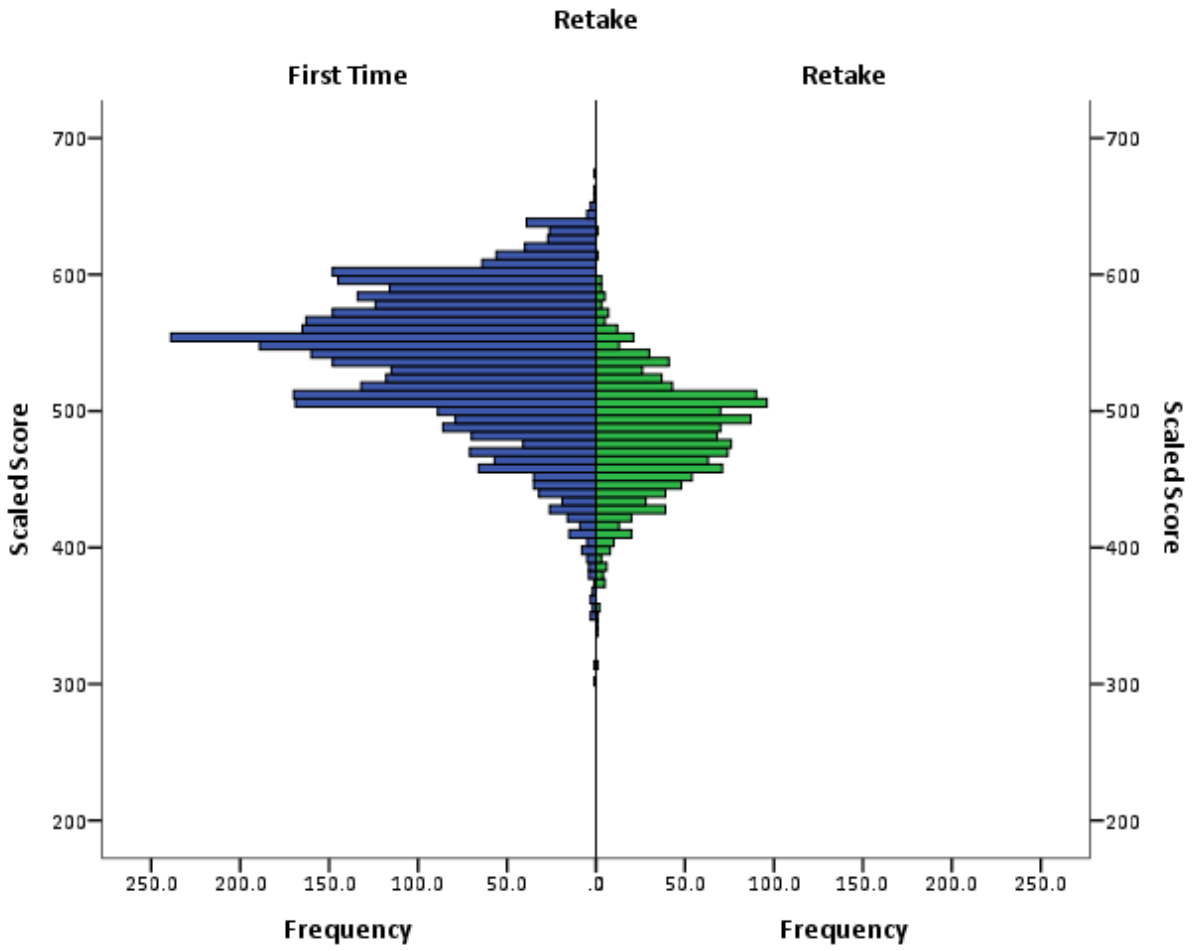
Figure 1: Cumulative Percentage of First-time and Retake Candidates by Scaled Score, BOC 2012-2013.



Ideally there should be a sharp increase in the cumulative proportion of candidates around the passing standard; that is, the slope of the curve would be more vertical around the passing standard. Test forms in which the slope is before or after the passing standard would not be functioning optimally. If the candidates were generally well-prepared for the certification examination, a relatively constrained set of scores with no long tails to the upper or lower end of the scale would be expected. The data in Figure 1 shows that the majority of candidates are performing consistently with this ideal. The figure also shows that first-time candidates are more successful in their performance than retake candidates.

Figure 2 provides information on the distribution of scale scores for the two cohorts of candidates.

Figure 2: Scale Score Distribution of First-time and Retake Candidates, BOC 2012-2013.



Test Form Summary Statistics

Table 4 provides test form descriptive statistics for each test window by form and retake status (see Appendix A for information on the statistics reported).

Table 4: Summary Test Form Statistics in Scaled Scores for Candidates for BOC Certification Examination, 2011-2012.

Test Window	Exam Form	Retake Status	N	Mean	Std. Dev.
April	362(7)	First Time	980	547	47
		Retake	98	485	44
		Total	*1,078*	*541*	*50*
	362(8)	First Time	989	546	48
		Retake	89	485	41
		Total	*1,078*	*541*	*50*
	Total	First Time	*1,969*	*547*	*48*
		Retake	*187*	*485*	*43*
		Total	*2,156*	*541*	*50*
June	362(9)	First Time	362	527	52
		Retake	175	490	41
		Total	*537*	*515*	*51*
	362(10)	First Time	359	528	52
		Retake	177	484	39
		Total	*536*	*514*	*53*
	Total	First Time	721	528	52
		Retake	352	487	40
		Total	*1,073*	*514*	*52*
August	362(11)	First Time	219	511	55
		Retake	305	482	41
		Total	*524*	*494*	*49*
	Total	First Time	219	511	55
		Retake	305	482	41
		Total	*524*	*494*	*49*
November	*362(12)*	First Time	251	527	54
		Retake	300	481	39
		Total	*551*	*502*	*52*
	Total	First Time	251	527	54
		Retake	300	481	39
		Total	*551*	*502*	*52*
February	*362(12)*	First Time	471	545	53
		Retake	175	482	42
		Total	*646*	*528*	*57*
	Total	First Time	471	545	53
		Retake	175	482	42
		Total	*646*	*528*	*57*
Form Totals	*362(7)*	First Time	980	547	47
		Retake	98	485	44
		Total	*1,078*	*541*	*50*
	362(8)	First Time	989	546	48
		Retake	89	485	41
		Total	*1,078*	*541*	*50*
	362(9)	First Time	362	527	52
		Retake	175	490	41
		Total	*537*	*515*	*51*
	362(10)	First Time	359	528	52
		Retake	177	484	39

Test Window	Exam Form	Retake Status	N	Mean	Std. Dev.
		Total	*536*	*514*	*53*
	362(11)	First Time	219	511	55
		Retake	305	482	41
		Total	*524*	*494*	*49*
	362(12)	First Time	722	539	54
		Retake	475	482	40
		Total	*1,197*	*516*	*56*
	Total	First Time	3,631	539	51
		Retake	1,319	484	41
		Total	*4,950*	*524*	*54*

As shown in Table 4, consistent with previous test administration years, there were differences in the scaled scores for each test window and by retake status. A Univariate General Linear Model (GLM) was conducted to assess the interaction between test form, test window, and retake status. The results indicated that there was no statistically significant performance difference by candidates on each test form, that there was no statistical interaction between retake status and test form (i.e., retake candidates performed the same across all test forms, as did first-time candidates), and that there were statistical differences in the performance of candidates for different test windows (month) and by retake status. Table 5 provides the Between-Subjects results for the Univariate GLM.

Table 5: Univariate Between Subject Effects Assessing Interaction Between Exam Form, Test Window, and Retake Status for BOC Certification Examination, 2012-2013.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Corrected Model	3434826.6	15	228988.439	100.486	0.000	0.234	1507.293	1.000
Intercept	72712952.0	1	72712952.050	31908.364	0.000	0.866	31908.364	1.000
Test Window	25083.1	1	25083.128	11.007	0.001	0.002	11.007	0.913
Retake	131839.3	1	131839.304	57.855	0.000	0.012	57.855	1.000
Exam Form	6761.0	3	2253.663	0.989	0.397	0.001	2.967	0.272
Test Window * Retake	18484.2	1	18484.208	8.111	0.004	0.002	8.111	0.813
Retake * Exam Form	4166.4	3	1388.811	0.609	0.609	0.000	1.828	0.178
Error	11255019.8	4939	2278.805					
Total	1376912216.0	4955						
Corrected Total	14689846.4	4954						

The candidates for the June 2012 test window had the highest scaled score, although their performance was not statistically different from April and February.

Difficulty and Discrimination

During the test administration year, item and test form performance are reviewed at every administration. For the annual summary, the item difficulty and discrimination statistics are reported for the first administration of the two scored sets administered. Data on the range of difficulty and discrimination statistics obtained for the first administration of the scored items is contained in Table 6 below.

Table 6: Summary of Item Discrimination and Difficulty for First Forms Tested, BOC 2012-2013.

Form		Difficulty		Discrimination	
Scored Set A	**Average**	**0.71**		**0.22**	
	Median	**0.74**		**0.23**	
	Minimum	**0.22**		**-0.05**	
	Maximum	**0.99**		**0.47**	
	Range	*Frequency*	*Percent*	*Frequency*	*Percent*
	< 0			1	0.6%
	0 to 0.1			16	9.1%
	0.1 to 0.2			48	27.4%
	0.2 to 0.3	5	2.9%	72	41.1%
	0.3 to 0.4	7	4.0%	31	17.7%
	0.4 to 0.5	9	5.1%	7	4.0%
	0.5 to 0.6	19	10.9%		
	0.6 to 0.7	27	15.4%		
	0.7 to 0.8	42	24.0%		
	0.8 to 0.9	47	26.9%		
	> 0.9	19	10.9%		
	Total		100%		100%
Scored Set B	**Average**	**0.67**		**0.22**	
	Median	**0.71**		**0.22**	
	Minimum	**0.13**		**-0.13**	
	Maximum	**0.98**		**0.56**	
	Range	*Frequency*	*Percent*	*Frequency*	*Percent*
	< 0			3	1.7%
	0 to 0.1			17	9.7%
	0.1 to 0.2	3	1.7%	49	28.0%
	0.2 to 0.3	6	3.4%	60	34.3%
	0.3 to 0.4	11	6.3%	38	21.7%
	0.4 to 0.5	11	6.3%	6	3.4%
	0.5 to 0.6	18	10.3%	2	1.1%
	0.6 to 0.7	32	18.3%		
	0.7 to 0.8	46	26.3%		
	0.8 to 0.9	35	20.0%		
	> 0.9	13	7.4%		
	Total		100%		100%

Overall, discrimination statistics for the items were acceptable, and average difficulty for the BOC certification examination forms was appropriate.

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 role delineation/practice analysis study (RD/PA6) introduced in April 2011. 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 RD/PA6.

Each test item has been linked to a specific content area of the test specifications, and items meet minimum standards of criticality for work as an entry-level athletic trainer. Thus, the procedures used to construct the BOC certification examination support the inference that the examination has been built to achieve its stated purpose. Consistent with the objectives of the BOC certification examination program, the examination 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 five performance domains in the content framework for the BOC examination, consistent with RD/PA6 upon which the certification examination is based. Table 7 reports descriptive statistics at the domain level using raw scores.

Table 7: Domain Level Statistics for Each Test Form for All Candidates for BOC Certification Examination, 2012-2013 (Raw Scores).

Form		N	Minimum	Maximum	Mean	Std. Dev.
362(7)	Prevention	1078	13.6	29.3	23.3	2.76
	Evaluation	1078	10.0	26.1	19.4	2.83
	Immediate Care	1078	11.4	24.3	19.6	2.11
	Treatment	1078	11.2	26.0	19.5	2.76
	Organization	1078	2.1	13.9	9.1	1.95
362(8)	Prevention	1078	11.9	29.7	23.5	2.83
	Evaluation	1078	10.8	26.3	19.5	2.81
	Immediate Care	1078	12.3	24.2	19.8	1.93
	Treatment	1078	9.7	25.8	19.5	2.74
	Organization	1078	3.3	13.7	8.9	1.94
362(9)	Prevention	537	11.8	29.0	21.9	3.11
	Evaluation	537	10.2	25.6	18.7	2.82
	Immediate Care	537	9.2	22.8	18.2	2.03
	Treatment	537	10.4	24.0	17.9	2.57
	Organization	537	2.5	13.8	8.3	1.96
362(10)	Prevention	536	11.0	29.5	21.7	3.13
	Evaluation	536	10.8	25.6	18.7	2.67
	Immediate Care	536	10.9	23.5	18.4	2.02
	Treatment	536	9.6	24.6	17.9	2.50
	Organization	536	2.1	13.7	8.2	2.08
362(11)	Prevention	524	11.3	29.8	21.2	2.95
	Evaluation	524	9.4	24.7	17.2	2.72
	Immediate Care	524	12.0	23.8	18.6	2.17
	Treatment	524	10.1	25.1	17.4	2.74
	Organization	524	2.8	13.0	7.9	1.91
362(12)	Prevention	1197	7.0	29.0	22.1	3.26
	Evaluation	1197	9.5	26.0	18.7	2.99
	Immediate Care	1197	9.2	23.5	18.4	2.08
	Treatment	1197	8.9	25.6	17.8	2.66
	Organization	1197	2.0	14.0	8.2	2.11

Correlations in candidate performance between the five domains ranged from 0.35 to 0.60, indicating that the domains were assessing somewhat different constructs (see Appendix B). These correlations are consistent with the results obtained for previous years.

Test Form Reliabilities & Other Summary Data

Data presented in Table 8 summarizes the performance of the test forms used for the BOC certification examination and is consistent with reporting requirements for NCCA/ICE accreditation. Reliability is assessed using Cronbach's alpha (Cronbach, 1951), a measure typically used for estimating reliability for tests that consist of non-binary data, and the standard error of measurement (SE*m*) presented in scaled score units.

Table 8: Summary Statistics for the 2012-2013 Administrations of BOC Athletic Trainer Test Forms.

Form #	Total No. of Candidates Tested	Percent Passing Form	Passing Point	Average Score	Standard Deviation	SE*m*	Reliability Estimate	Total No. of Items on Form
362(7)	1,078	81%	500	541	50	18.04	0.87	175
362(8)	1,078	82%	500	541	50	18.47	0.86	175
362(9)	537	65%	500	515	51	19.91	0.85	175
362(10)	536	63%	500	514	53	19.72	0.86	175
362(11)	524	49%	500	494	49	19.69	0.84	175
362(12)	1,197	62%	500	516	56	18.05	0.90	175
Total	4,950	70%		524	52	18.71	0.87	

Data presented Table 8 is in scaled score units for passing point, average score, standard deviation, and standard error of measurement.

Data presented in Table 8 shows that each test window meets general guidelines for a reliabilities greater than 0.70 and is consistent with previous years. Standard errors of measurement also are consistent with previous years.

SUMMARY

Statistics concerning the quality of the BOC certification examination 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.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

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 can be placed 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 one 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 0.30 (difficult) to 0.92 (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., 0.25 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, having more items on a test increases the test's reliability.

Reliability does not imply validity. A reliable measure is measuring something consistently, but the statistics does not specify 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 Five Domains

Correlations

		DOMAIN_01	DOMAIN_02	DOMAIN_03	DOMAIN_04	DOMAIN_05
DOMAIN_01	Pearson Correlation	1	.603**	.467**	.555**	.476**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	4955	4955	4955	4955	4955
DOMAIN_02	Pearson Correlation	.603**	1	.453**	.578**	.448**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	4955	4955	4955	4955	4955
DOMAIN_03	Pearson Correlation	.467**	.453**	1	.435**	.353**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	4955	4955	4955	4955	4955
DOMAIN_04	Pearson Correlation	.555**	.578**	.435**	1	.430**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	4955	4955	4955	4955	4955
DOMAIN_05	Pearson Correlation	.476**	.448**	.353**	.430**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	4955	4955	4955	4955	4955

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