



**BOARD OF CERTIFICATION  
FOR THE ATHLETIC TRAINER**

*Be Certain.™*

**Examination Report for 2011-2012 Testing Year**

**Board of Certification (BOC) Certification Examination for Athletic Trainers**

Stephen B. Johnson, Ph.D.  
Castle Worldwide, Inc.  
Prepared March 2012  
Updated May 2012

## TABLE OF CONTENTS

INTRODUCTION .....	3
Description of the Certification Examination.....	3
Format of Items on the Certification Examination .....	4
Delivery of the Certification Examination .....	4
Number of Test Forms .....	5
Standard Setting and Equating of Test Forms .....	5
Use of Scaled Scores .....	5
Score Reporting.....	5
Certification Examination Development.....	6
Stand-Alone Committee .....	6
Focused Testlet Committee .....	6
Test Form Assembly.....	6
I-Dev.....	6
Item Bank.....	6
ANALYSIS OF THE CERTIFICATION EXAMINATION .....	7
Candidate Performance .....	7
Candidates Excluded from this Report .....	7
Pass Rates .....	7
Distribution of Candidate Scores .....	9
Test Form Summary Statistics.....	11
Difficulty and Discrimination.....	13
Domain Performance.....	14
Test Form Reliabilities & Other Summary Data .....	16
SUMMARY .....	17
REFERENCES .....	18
APPENDICES .....	19
Appendix A: Definitions of Form Statistics .....	19
Mean Score.....	19
Standard Deviation.....	19
Standard Error of Measurement .....	19
Min and Max (Low and High Score) .....	19
Avg. Diff .....	19
Avg. Discrim.....	20
Reliability Measures .....	20
Appendix B: Correlations of Candidate Performance on the Five Domains.....	21

**FIGURES**

Figure 1: Cumulative Percentage of First-time New Graduates and Retake Candidates by Scaled Score, BOC 2011-2012. .... 9

Figure 2: Scale Score Distribution of First-time New Graduates and Retake Candidates,BOC 2011-2012. .... 10

**TABLES**

Table 1: Number of Candidates in Three Cohorts and Pass Rates for BOC Certification Examination, 2005-2006 to 2011-2012. .... 8

Table 2: Passing Rates for Each Test Form for All Candidates for BOC Certification Examination, 2011-2012..... 8

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, 2010-2011. .... 9

Table 4: Summary Test Form Statistics in Scaled Scores for Candidates for BOC Certification Examination, 2011-2012. .... 11

Table 5: Univariate Between Subject Effects Assessing Interaction Between Exam Form (ExamForm), Test Window (Month), and Retake Status (Retake) for BOC Certification Examination, 2011-2012. .... 13

Table 6: Summary of Item Discrimination and Difficulty for First Forms Tested, BOC 2011-2012..... 14

Table 7: Domain Level Statistics for Each Test Form for All Candidates for BOC Certification Examination, 2010-2011 (Raw Scores). .... 15

Table 8: Summary Statistics for the 2011-2012 Administrations of BOC Athletic Trainer Test Forms..... 16

## INTRODUCTION

The Board of Certification (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 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 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. The certification examination blueprint is contained in Appendix A.

### ***Format of Items on the Certification Examination***

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

The multiple-choice items contain a stem and four to 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” in to one of the “buckets.” Options can be used once, multiple times, or not at all 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 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/](http://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 a combination of scored and experimental [unscored] items totaling 175 total items. Each test form includes five focused testlets.

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 2011-2012 testing year, the certification examination was administered in five 14-day test windows: March/April 2011, May/June 2011, July/August 2011, November 2011, and January/February 2012. 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***

Multiple sets scored items were developed for 2011-2012. Each scored set was assigned different experimental sets for the year, creating six different test forms. Forms 362(1) and 362(2) were administered in April 2011, Forms 362(3) and 362(4) in June 2011, Form 362(5) in August 2011, and Form 362(6) in November 2011 and February 2012.

### ***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) scheduled for introduction in April 2011. The panel participated in three rounds of data collection using a modified-Angoff model, the Yes/No technique (Impara & Plake, 1997).

Forms 362(1) and 362(2) contained the same 125 scored items presented to the standard setting panel, with a different set of experimental items. Form 362(1) was administered to 1,031 candidates, and Form 362(2) was administered to 977 candidates. Candidates were randomly assigned to each form.

Following administration, a review of the standard setting panel's data and performance of candidates on both forms were presented to the BOC. The BOC confirmed the use of the panel's median recommendation from the third round of review.

Forms 362(3) and 362(4) administered in June 2011 had items in common with the scored set of items used for the standard setting form. 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 standard setting form, 362(1).

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 this period, the certification examination consists of one assessment experience for candidates. During the 2008-2009 testing year, focused testlets were introduced to the testing model.

### ***Stand-Alone Committee***

During the February 2011 meeting, the committee developed and reviewed 66 stand-alone items. During the July 2011 meeting, the committee reviewed 107 items that were field tested in April and June 2011. In addition, the committee developed and reviewed 26 stand-alone items. During the November 2011 meeting, the committee reviewed and finalized 225 items in order to assemble two forms of the BOC certification examination.

### ***Focused Testlet Committee***

During the February 2011 meeting, the committee reviewed and finalized two focused testlets and began development of 11 focused testlets. In addition, one focused testlet was broken into five stand-alone items. During the July 2011 meeting, the committee reviewed eight focused testlets that were field tested in April and June 2011. In addition, five focused testlets were reviewed and finalized, and six focused testlets were developed. During the November 2011 meeting, the committee reviewed and finalized 20 stand-alone items and eight focused testlets.

### ***Test Form Assembly***

As part of RD/PA6, a new test blueprint was developed and approved. During the November 2010 meeting, two sets of scored items aligned to the new test blueprint were assembled for administration during the 2011-2012 testing year. A second review of the standard setting form was conducted in January 2011 in preparation for the standard setting meeting in February 2011.

### ***I-Dev***

The BOC also uses Castle's I-Dev system for online development of multiple-choice items. In 2011, BOC subject matter experts completed development of 608 items. The review and validation of these items will continue into 2012.

### ***Item Bank***

Currently, BOC's item bank includes multiple-choice, multi-select, drag-and-drop, and hotspot items. Castle staff continually reviews and edits items and the resulting certification examination forms for psychometric and publication purposes. Items for the certification examination are stored in I-Bank, Castle's proprietary item-banking system.

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

There were 4,886 reported administrations of the BOC certification examination during the 2011-2012 testing year, a decline of 14% from 2010-2011 (5,711), and a 20% decline from 2009-2010 (6,171) and 2008-2009 (6,135). Continuing an upward trend since 2008-2009, of the 4,886 administrations, 3,222 (66%) examinations were administered to first-time candidates, compared with 52% in 2010-2011 and 46% in 2008-2009 and 2009-2010.

### ***Pass Rates***

Table 1 provides annual pass rates for 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.<sup>1</sup>

Year	First-time	Pass	% Pass	Retake	Pass	% Pass	All	Pass	% Pass
<b>RD5</b>									
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%
<b>RD6</b>									
2011-2012	3,222	2,653	82.3%	1,664	696	41.8%	4,886	3,269	66.9%

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, 2011-2012.

Test Window	Form	Frequency			Percent	
		Fail	Pass	Total	Fail	Pass
April 2011	362(1)	234	797	1031	22.7%	77.3%
	362(2)	234	743	977	24.0%	76.0%
	<i>Total</i>	<i>468</i>	<i>1540</i>	<i>2008</i>	<i>23.3%</i>	<i>76.7%</i>
June 2011	362(3)	198	367	565	35.0%	65.0%
	362(4)	220	392	612	35.9%	64.1%
	<i>Total</i>	<i>418</i>	<i>759</i>	<i>1177</i>	<i>35.5%</i>	<i>64.5%</i>
August 2011	362(5)	300	286	586	51.2%	48.8%
	362(1)	1	1	2	50.0%	50.0%
	<i>Total</i>	<i>301</i>	<i>287</i>	<i>588</i>	<i>51.2%</i>	<i>48.8%</i>
November 2011	362(6)	263	291	554	47.5%	52.5%
	362(1)	1	0	1	100.0%	0.0%
	<i>Total</i>	<i>264</i>	<i>291</i>	<i>555</i>	<i>47.6%</i>	<i>52.4%</i>
February 2012	362(6)	166	392	558	29.7%	70.3%
	<i>Total</i>	<i>166</i>	<i>392</i>	<i>558</i>	<i>29.7%</i>	<i>70.3%</i>
<b>ALL</b>		<b>1617</b>	<b>3269</b>	<b>4886</b>	<b>33.1%</b>	<b>66.9%</b>

<sup>1</sup> 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 2011-2012 with a comparison of the performance of candidates since 2008-2009.

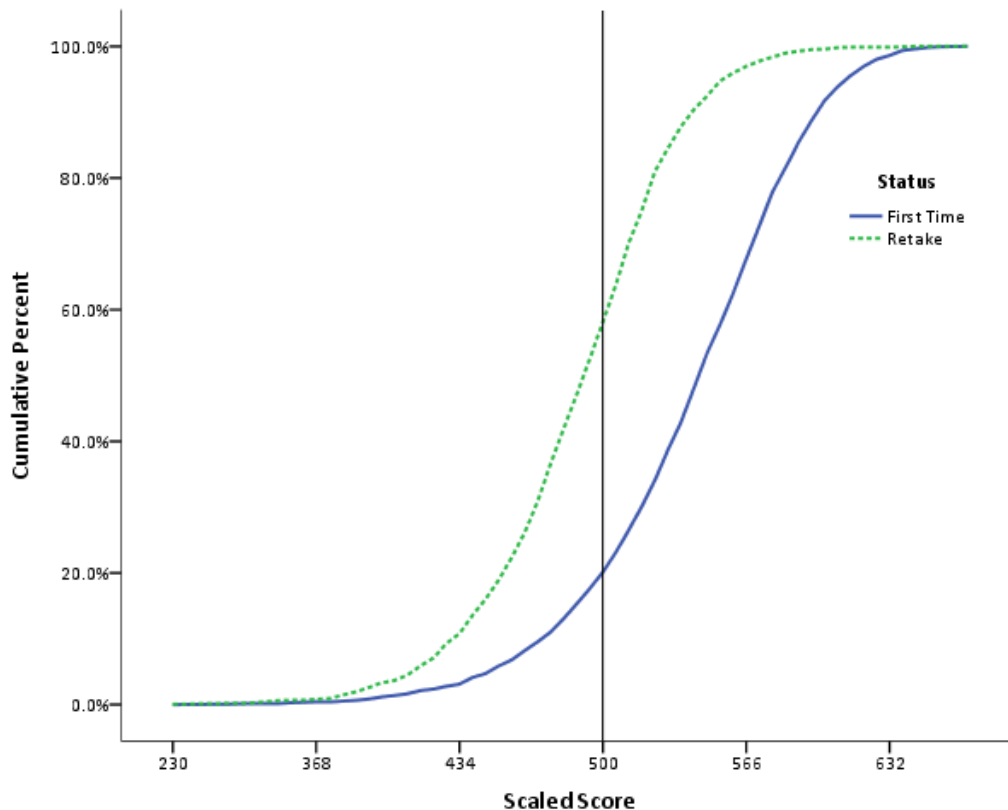
**Table 3:** Number of Candidates in Three Cohorts, Minimum, Maximum and Average Scaled Score, Median and Mode Scaled Score, and Standard Deviation (Scaled Score) for BOC Certification Examination, 2010-2011.

Cohort	N	Avg.	Median	Std. Dev.	Min	Max
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

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 ( $F(1, 4885) = 1195.59, p < .001, \eta = .197$ ).

Figure 1 presents a cumulative frequency distribution for 2011-2012 retake and first-time candidates. The figure represents the proportion of candidates who scored at a scale score or lower.

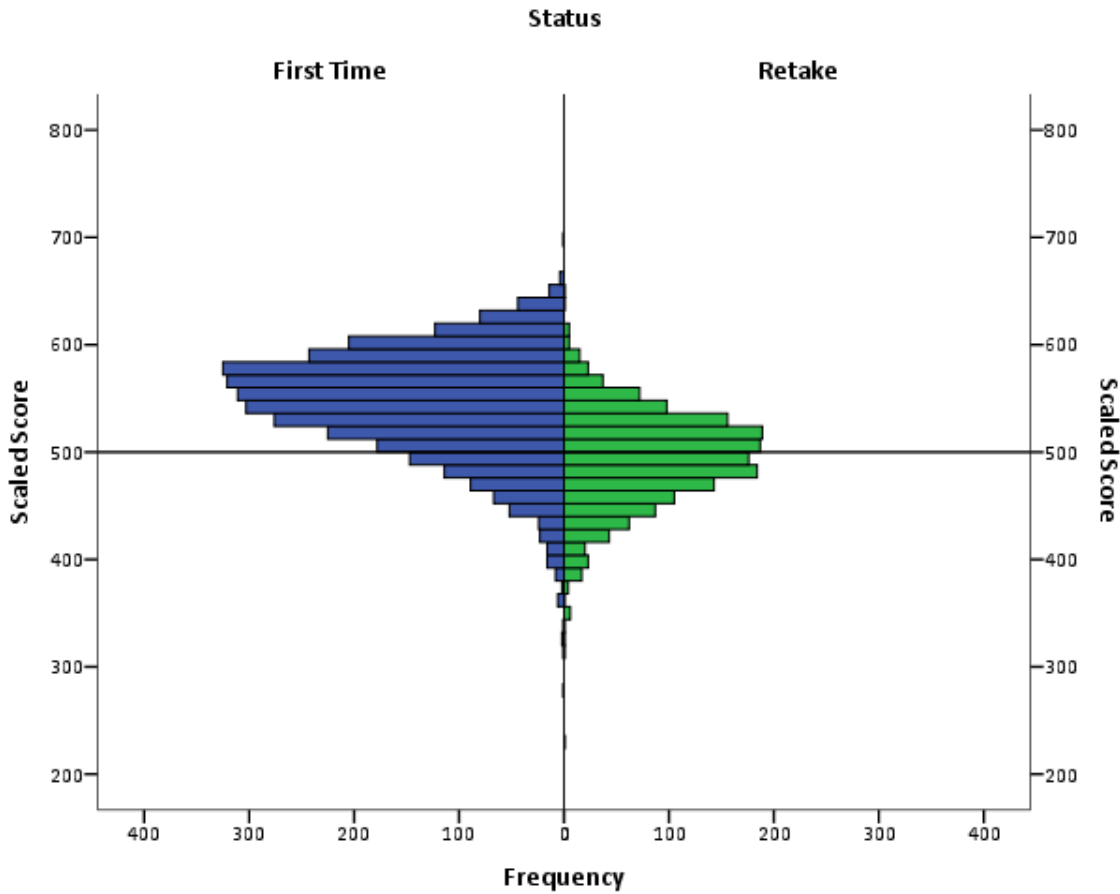
**Figure 1:** Cumulative Percentage of First-time and Retake Candidates by Scaled Score, BOC 2011-2012.



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 where 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, we also would expect to see a relatively constrained set of scores, no long tails to the upper or lower end of the scale. The data in Figure 1 shows that the majority of candidates are performing consistent 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 2011-2012.



### 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 2011	362(1)	First-time	832	550	48
		Retake	199	497	48
		Total	1031	539	52
	362(2)	First-time	774	547	47
		Retake	203	495	46
		Total	977	536	52
	<i>Total</i>	First-time	1606	548	48
		Retake	402	496	47
		Total	2008	538	52
June 2011	362(3)	First-time	373	534	51
		Retake	192	492	42
		Total	565	519	52
	362(4)	First-time	397	534	54
		Retake	215	496	46
		Total	612	521	54
	<i>Total</i>	First-time	770	534	53
		Retake	407	494	44
		Total	1177	520	53
August 2011	362(1)	Retake	2	452	76
		Total	2	452	76
	362(5)	First-time	238	521	49
		Retake	348	486	43
		Total	586	501	48
	<i>Total</i>	First-time	238	521	49
		Retake	350	486	43
		Total	588	500	49
	November 2011	362(1)	First-time	1	494
Total			1	494	.
362(6)		First-time	256	533	59
		Retake	298	488	42
		Total	554	509	56
<i>Total</i>		First-time	257	533	59
		Retake	298	488	42
		Total	555	509	56
February 2012		362(6)	First-time	351	551
	Retake		207	489	44
	Total		558	528	55
	<i>Total</i>	First-time	351	551	47

Test Window	Exam Form	Retake Status	N	Mean	Std. Dev.
		Retake	207	489	44
		Total	558	528	55
Form Totals for 2011-2012	362(1)	First-time	833	549	48
		Retake	201	497	48
		Total	1034	539	52
	362(2)	First-time	774	547	47
		Retake	203	495	46
		Total	977	536	52
	362(3)	First-time	373	534	51
		Retake	192	492	42
		Total	565	519	52
	362(4)	First-time	397	534	54
		Retake	215	496	46
		Total	612	521	54
	362(5)	First-time	238	521	49
		Retake	348	486	43
		Total	586	501	48
	362(6)	First-time	607	544	53
		Retake	505	488	43
		Total	1112	518	56
	<i>Total</i>	First-time	3222	542	51
		Retake	1664	491	44
		Total	4886	525	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 (ExamForm), Test Window (Month), and Retake Status (Retake) for BOC Certification Examination, 2011-2012.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Corrected Model	3136135.06	15	209075.7	90.448	.000	.218	1356.720	1.000
Intercept	33049897.00	1	33049897.0	14297.676	.000	.746	14297.676	1.000
ExamForm	6831.11	4	1707.8	.739	.565	.001	2.955	.241
Month	33809.67	3	11269.9	4.875	.002	.003	14.626	.909
Retake	2074815.57	1	2074815.6	897.583	.000	.156	897.583	1.000
ExamForm * Month	.00	0				.000	.000	
ExamForm * Retake	815.59	2	407.8	.176	.838	.000	.353	.077
Month * Retake	20237.39	1	20237.4	8.755	.003	.002	8.755	.841
ExamForm * Month * Retake	.00	0				.000	.000	
Error	11257284.14	4870	2311.6					
Total	1359344240.00	4886						
Corrected Total	14393419.21	4885						

In general, consistent with previous years, the performance of candidates in April 2011 and February 2012 is higher than other periods of the year, with most of the difference accounted for by the performance of retake candidates.

***Difficulty and Discrimination***

During the test administration year, item and test form performance is 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, Forms 362(1) and 362(3). Data on the range of difficulty and discrimination statistics obtained for the 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 2011-2012.

<b>Form</b>	<b>Difficulty</b>			<b>Discrimination</b>	
362(1)	<b>Average</b>	<b>0.72</b>		<b>0.20</b>	
	<b>Median</b>	<b>0.74</b>		<b>0.21</b>	
	<b>Minimum</b>	<b>0.01</b>		<b>-0.18</b>	
	<b>Maximum</b>	<b>1.00</b>		<b>0.46</b>	
	<i>Range</i>	<i>Frequency</i>	<i>Percent</i>	<i>Frequency</i>	<i>Percent</i>
	<0			4	3.2%
	0 to .1	2	1.6%	20	16.0%
	.1 to .2	0	0.0%	36	28.8%
	.2 to .3	5	4.0%	41	32.8%
	.3 to .4	2	1.6%	20	16.0%
	.4 to .5	9	7.2%	4	3.2%
	.5 to .6	10	8.0%		
	.6 to .7	21	16.8%		
	.7 to .8	25	20.0%		
	.8 to .9	30	24.0%		
	>.9	21	16.8%		
	<i>Total</i>	125	100%	125	100%
362(3)	<b>Average</b>	<b>0.70</b>		<b>0.20</b>	
	<b>Median</b>	<b>0.74</b>		<b>0.20</b>	
	<b>Minimum</b>	<b>0.01</b>		<b>-0.18</b>	
	<b>Maximum</b>	<b>1.00</b>		<b>0.50</b>	
	<i>Range</i>	<i>Frequency</i>	<i>Percent</i>	<i>Frequency</i>	<i>Percent</i>
	<0			4	3.2%
	0 to .1	2	1.6%	19	15.2%
	.1 to .2	1	0.8%	41	32.8%
	.2 to .3	4	3.2%	36	28.8%
	.3 to .4	3	2.4%	20	16.0%
	.4 to .5	7	5.6%	4	3.2%
	.5 to .6	12	9.6%	1	0.8%
	.6 to .7	23	18.4%		
	.7 to .8	29	23.2%		
	.8 to .9	28	22.4%		
	>.9	16	12.8%		
	<i>Total</i>	125	100%	125	100%

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.

### **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 to 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, 2010-2011 (Raw Scores).

Form		N	Minimum	Maximum	Mean	Std. Dev.
362(1)	Prevention	1034	10.7	28.5	22.1	2.74
	Evaluation	1034	10.0	27.2	20.3	2.80
	Immediate Care	1034	10.3	23.6	18.7	1.97
	Treatment	1034	10.2	25.5	19.4	2.34
	Organization	1034	3.2	15.0	9.1	2.11
362(2)	Prevention	977	11.6	28.1	21.9	2.67
	Evaluation	977	9.8	26.6	20.2	2.74
	Immediate Care	977	8.8	23.3	18.6	1.96
	Treatment	977	10.7	25.5	19.4	2.26
	Organization	977	2.4	14.6	8.9	2.14
362(3)	Prevention	565	11.0	27.7	21.2	2.80
	Evaluation	565	11.0	26.4	19.5	2.91
	Immediate Care	565	10.4	22.4	18.3	1.82
	Treatment	565	11.2	24.3	18.8	2.33
	Organization	565	2.2	14.7	8.4	2.01
362(4)	Prevention	612	11.5	28.1	21.2	2.68
	Evaluation	612	10.8	26.7	19.6	2.89
	Immediate Care	612	11.6	22.8	18.1	1.98
	Treatment	612	10.7	24.9	18.9	2.44
	Organization	612	2.0	14.0	8.6	2.13
362(5)	Prevention	586	10.2	27.0	20.5	2.70
	Evaluation	586	10.7	26.1	18.7	2.61
	Immediate Care	586	5.0	22.2	17.7	1.99
	Treatment	586	6.6	25.0	18.0	2.41
	Organization	586	1.8	14.7	8.1	1.98
362(6)	Prevention	1112	11.9	27.9	21.0	2.82
	Evaluation	1112	8.1	26.4	19.4	3.03
	Immediate Care	1112	8.1	23.1	18.2	2.02
	Treatment	1112	7.1	25.3	18.8	2.43
	Organization	1112	2.2	14.7	8.6	2.13



Correlations in candidate performance between the five domains ranged from 0.327 to 0.562, 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 the Table 8 summarizes the performance of the test forms used for the BOC certification examination and are consistent with reporting requirements for NCCA/ICE accreditation. Reliability is assessed using the 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 (*SEm*) presented in scaled score units.

**Table 8:** Summary Statistics for the 2011-2012 Administrations of BOC Athletic Trainer Test Forms.

Form #	Total No. of Candidates Tested	Percent Passing Form	Passing Point	Average Score	Standard Deviation	<i>SEm</i>	Reliability Estimate	Total No. of Items on Form
362(1)	1034	77%	500	539	52	19.08	0.87	175
362(2)	977	76%	500	536	52	19.72	0.85	175
362(3)	565	65%	500	519	52	19.07	0.87	175
362(4)	612	64%	500	521	54	20.22	0.86	175
362(5)	586	49%	500	501	48	19.97	0.83	175
362(6)	1112	61%	500	518	55	21.90	0.85	175
Total	4886	66.91%		525	53	20.11	0.85	

Data presented Table 8 is in scale 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 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 ( $SEm$ ) is used to determine the range of certainty around a candidate's reported score. The  $SEm$  makes it possible to determine how reliable a particular test is and how much confidence we can place in the scores it yields.

The  $SEm$  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  $SEm$  is used to create upper and lower boundaries around an observed score. The lower the  $SEm$  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., .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 Five Domains**

		<b>Correlations</b>				
		DOMAIN_01	DOMAIN_02	DOMAIN_03	DOMAIN_04	DOMAIN_05
DOMAIN_01	Pearson Correlation	1	0.562**	0.398**	0.473**	0.418**
	Sig. (2-tailed)		0.000	0.000	0.000	0.000
	N	4886	4886	4886	4886	4886
DOMAIN_02	Pearson Correlation	0.562**	1	0.411**	0.521**	0.402**
	Sig. (2-tailed)	0.000		0.000	0.000	0.000
	N	4886	4886	4886	4886	4886
DOMAIN_03	Pearson Correlation	0.398**	0.411**	1	0.360**	0.327**
	Sig. (2-tailed)	0.000	0.000		0.000	0.000
	N	4886	4886	4886	4886	4886
DOMAIN_04	Pearson Correlation	0.473**	0.521**	0.360**	1	0.361**
	Sig. (2-tailed)	0.000	0.000	0.000		0.000
	N	4886	4886	4886	4886	4886
DOMAIN_05	Pearson Correlation	0.418**	0.402**	0.327**	0.361**	1
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	
	N	4886	4886	4886	4886	4886

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