

# **The Connecticut Academic Performance Test: Technical Report**

**Prepared by  
Irene Hendrawan & Arianto Wibowo**

**January 2013**



# Table of Contents

<b>Table of Contents</b> .....	i
<b>List of Charts</b> .....	ii
<b>List of Tables</b> .....	iii
<b>Part 1: Introduction</b> .....	1
1.1. General description of CAPT.....	1
1.2. 2012 CAPT Test Design.....	1
1.3. 2012 CAPT Test Forms.....	2
<b>Part 2: Test Development</b> .....	3
<b>Part 3: Item Level Statistics</b> .....	4
<b>Part 4: Scaling and Equating</b> .....	5
4.1. 2012 CAPT Linking Items.....	5
4.2. Calibration Process.....	6
<b>Part 5: Test Statistics</b> .....	9
5.1. Reliability.....	9
5.2. Classification Consistency and Accuracy.....	9
<b>Part 6: CAPT3 Standards</b> .....	11
<b>Part 7: Validity</b> .....	13
7.1. Content Validity Survey.....	13
7.2. Scoring Quality Assurance Procedures Undertaken during Development.....	13
7.3. Item Quality Analysis Undertaken During Development.....	13
7.4. Equating Design.....	15
<b>References</b> .....	16
<b>Appendix A: Item Analysis</b> .....	17
<b>Appendix B: Raw, Theta, and Scale Scores</b> .....	23

## List of Charts

<b>Chart 1: Calibration Design for 2012 Mathematics</b> .....	7
<b>Chart 2: Calibration Design for 2012 Science</b> .....	7
<b>Chart 3: Calibration Design for 2012 Reading</b> .....	7
<b>Chart 4: Calibration Design for 2012 Writing</b> .....	7

## List of Tables

<b>Table 1: 2012 CAPT Operational Test Design</b> .....	1
<b>Table 2: Summary of Item Analysis Form HS19</b> .....	4
<b>Table 3: 2012 Embedded Linking Items</b> .....	5
<b>Table 4: 2012 CAPT Equating Constants</b> .....	7
<b>Table 5: Summary of Weighting for Reading and Writing</b> .....	8
<b>Table 6: Scaling Coefficients from Base Year (CAPT2)</b> .....	8
<b>Table 7: CAPT Cronbach’s Alpha</b> .....	9
<b>Table 8: CAPT Scale Score Summary Statistics</b> .....	9
<b>Table 9: Classification Consistency</b> .....	9
<b>Table 10: Classification Accuracy</b> .....	10
<b>Table 11: False Negative Classification</b> .....	10
<b>Table 12: False Positive Classification</b> .....	10
<b>Table 13: 2012 CAPT Achievement Levels and Scale Score Ranges</b> .....	12

# Part 1: Introduction

## 1.1. General description of CAPT

The Connecticut Academic Performance Test (CAPT) was designed to measure student performance in high school. Students are tested in the areas of Mathematics, Science, Reading, and Writing.

The CAPT has measured achievement of Connecticut students since 1994, when it was first administered. The second generation of CAPT was introduced in 2001. The content structure of the first generation CAPT was used as the baseline in developing the second generation. For the most part, the educational outcomes tested in the first generation were carried over to the second generation. Changes were made in light of new trends in instruction, educational assessment, and the lessons learned over the years of the first generation. The third generation of CAPT was introduced in the spring of 2007. The spring 2012 administration was the sixth operational (OP) administration of CAPT3.

## 1.2. 2012 CAPT Test Design

The spring 2012 administration comprises the following content areas:

1. Mathematics  
Mathematics (MA) has thirty-two operational items -- twenty-four grid-in (GR) response items and eight open-ended (OE) items scored on 0-3 scale.
2. Science  
Science (SC) has sixty-five OP items -- sixty multiple choice (MC) items and five OE items scored on 0-3 scale.
3. Reading  
Reading (RD) consists of two subtests:
  - Reading for Information  
Reading for Information (RI) has eighteen OP items -- twelve MC items and six OE items scored on 0-2 scale.
  - Response to Literature  
Response to Literature (RL) consists of an extended response (EX) item with a 2-12 score scale (sum of two rater scores on a 1-6 scale).
4. Writing  
Writing (WR) consists of three subtests:
  - Editing & Revising  
Editing & Revising (ER) has eighteen MC items.
  - Interdisciplinary Writing 1 & Interdisciplinary Writing 2  
Interdisciplinary Writing 1 (IW1) & Interdisciplinary Writing 2 (IW2) have EX items with a 2-12 score scale (sum of two rater scores on a 1-6 scale).

**Table 1: 2012 CAPT Operational Test Design**

Content Area	Subject	Number of Items				Total Items	Raw Score
		MC	GR	OE	EX		
Mathematics	Mathematics		24	8		32	0 - 48
Science	Science	60		5		65	0 - 75
Reading	Reading for Information	12		6		18	0 - 24
	Response to Literature				1	1	2 - 12
Writing	Editing & Revising	18				18	0 - 18
	Interdisciplinary Writing 1				1	1	2 - 12
	Interdisciplinary Writing 2				1	1	2 - 12

### **1.3. 2012 CAPT Test Forms**

In the 2012 administration, two main forms were available for administration: Form HS19, which is the live form taken by most of the students, and Form HS0, which was available for breach situations. Form HS0 will be used as a breach form throughout the third generation. Although the two forms were pre-equated during test assembly, there was still a need to carry out a post equating procedure after the test administration in order to ensure the comparability of the two forms.

In addition, CSDE piloted new items in 2013 forms. The pilot forms were administered to schools stratified by last year's achievement scores. CSDE's rationale for stratifying the test forms based on scale scores from the previous year was that this procedure would more likely yield groups of test takers who were representative with respect to the distribution of skills and achievement across the entire state. In other words, instead of sampling based on conventional demographic variables to achieve representation of test-taker characteristics, CSDE chose to sample on test-taker achievement. MI selects a stratified sample of schools, based on the scale score distribution to which each belongs.

Any student who breaches a test session or subtest (HS19 or HS0) was given the corresponding test session or subtest (HS19 or HS0).

## Part 2: Test Development

The process by which each form of the CAPT is developed is extensive, spanning a five- or six-year period and many stages. The development process is led and overseen by staff members in the Bureau of Student Assessment at the Connecticut State Department of Education (CSDE), but it also involves many other people who represent a wide variety of perspectives and areas of expertise. CSDE curriculum specialists and content experts play a critical role and work closely with the assessment staff throughout the process. In addition, a major testing company and other organizations and individuals with experience in educational assessment are involved at appropriate points in the development process.

Advisory committees of Connecticut educators are particularly important throughout the development of the CAPT. Content Advisory and Fairness Committees review each item to ensure the match between the content objectives and the items, and to ensure meaningful interpretability of test results. The Content Advisory Committees included content experts, regular and special education teachers, Connecticut State Department of Education curriculum, and assessment content specialists. A separate advisory committee is established for each part of the CAPT: Mathematics, Science, Reading, and Writing. These advisory committee members are selected on the basis of their knowledge in educational content and processes. In addition, the Fairness Committee is responsible for determining whether items are appropriate and fair to all examinees. Educators are carefully selected for the advisory committees to be representative of school districts throughout Connecticut.

The test development process for CAPT3 began with content specialists and testing experts writing test specifications with the help of the CAPT content advisory committees. The starting point for this process was looking at the specifications and structure of the first generation CAPT, and examining what has been working and what needed improvement. The new curriculum frameworks adopted by the State of Connecticut were also used as a guide. Test items for the CAPT3 were carefully developed in accordance with the established test specifications and test blueprint. These items were carefully matched to the content standards in the Connecticut Curriculum Frameworks for Mathematics, Science, Reading, and Writing. Items that did not pass the scrutiny of either Content Advisory or Fairness Committees were eliminated from the pool of pilot items.

After committee reviews, field test forms were created and piloted on a representative sample, stratified by scale score distribution, consisting of approximately 2000 students per form. Pilot statistics such as the mean, point biserial, and Rasch difficulty were generated and reviewed by CSDE assessment content staff and psychometricians. In addition, for hand-scored constructed response items, the contractor staff provided qualitative summaries about whether students appeared to have sufficient contextual knowledge to be able to fully respond to the item. Flawed items were removed from the item pool, including those showing test item bias or inappropriate levels of difficulty. Based on the CAPT3 Blueprints, Mathematics, Science, Reading, and Writing test forms of equivalent difficulty were simultaneously constructed from the pool of items that met all the review criteria. Every effort was made to ensure that strand level difficulties were comparable and that the items reflected the appropriate range of content within the strands across the generation.

## Part 3: Item Level Statistics

Table 2 and Appendix A present a summary and detailed results of item analysis (item quality) data, respectively for Mathematics, Science, Reading and Writing. The following information is presented in each item analysis:

**Classical and IRT difficulties:** Item difficulty is fundamentally a ratio of the proportion of examinees who answered the item correctly. Thus, an easy item has a high p-value and a difficult item has a low p-value. If an item has a very high p-value it may be so easy that it does not provide much information about what most examinees know or can do, while an item with a very low p-value may be so difficult that it is beyond the range of what most students know or can do. Therefore, items with very high or very low p-values may be rejected, unless content relevance overrides that concern.

The IRT difficulty described here is the Rasch IRT model's item difficulty parameter. This parameter influences the probability of correctly responding to the item as defined by the Rasch IRT model. For a given examinee's ability, the higher the IRT difficulty, the lower the probability of responding correctly. Thus, an easy item has a low Rasch difficulty and a difficult item has a high Rasch difficulty.

**Item Discriminations:** The point biserial correlation or item-total correlations measure the strength of the relationship between the particular item score and the total test score. Thus, item discrimination reflects how well a particular item differentiates between high and low total test performers. When the correlation is high, examinees that do well on the item also tend to do well on the entire test and correspondingly, examinees that do not do well on the item also tend not to do well on the total test.

**Distractor Frequencies:** The proportion of students who answered each option (A-D, 0-3, and 2-12) are presented for the multiple-choice items, open-ended and extended response, respectively.

**Table 2: Summary of Item Analysis Form HS19**

Subject	Rasch		P-value		Point Biserial	
	Mean	Std	Mean	Std	Mean	Std
Mathematics	-0.0894	0.8021	0.75	0.45	0.55	0.10
Reading for Information	0.0964	0.8197	0.69	0.13	0.41	0.12
Response to Literature	0.5998	-	7.05	-	0.63	-
Editing and Revising	0.0469	0.6106	0.75	0.10	0.32	0.06
Interdisciplinary Writing	1.4214	0.0139	7.64	0.04	0.73	0.01
Science	0.0545	0.6336	0.70	0.36	0.41	0.10

## Part 4: Scaling and Equating

### 4.1 2012 CAPT Linking Items

The 2012 CAPT Mathematics, Science, Reading for Information, and Editing & Revising tests were equated with the 2011 CAPT (HS18) subtests by embedding linking items. Linking items were counted toward students' scores.

The Live form of the 2012 CAPT (HS19) included:

- Mathematics – twelve linking grid items were embedded.
- Science – fifteen linking MC items were embedded.
- Editing & Revising – one passage with six linking MC items were embedded.
- Reading for Information – one passage with four linking MC items and two linking OE items.

Table 3 indicates the linking items used as well as their positions on the 2012 and 2011 tests.

**Table 3: 2012 Embedded Linking Items**

Content Area	Form HS19 Item Position	Form HS18 Item Position	Rasch Form HS18
Mathematics	5	5	-1.2992
	6	6	0.5157
	7	7	-0.0819
	8	8	0.8834
	15	15	1.1557
	16	16	0.2391
	23	23	-0.7974
	25	25	-0.1335
	26	26	-0.5483
	28	28	0.2444
	30	30	0.5580
	32	32	-0.2424
Science	5	5	-0.7487
	8	8	0.6008
	9	9	-0.9867
	10	10	-0.6259
	11	11	-0.5242
	22	22	0.6395
	31	31	-0.3209
	35	35	0.4364
	36	36	-0.0404

Content Area	Form HS19 Item Position	Form HS18 Item Position	Rasch Form HS18
	37	37	-0.7038
	38	38	-0.4456
	44	44	-0.3281
	45	45	1.2958
	46	46	-0.7592
	49	49	0.7268
Reading	7	7	-0.0351
	8	8	-0.6991
	9	9	0.0589
	10	10	-0.6438
	11	11	1.0177
	12	12	0.9372
Writing	1	1	-0.7653
	2	2	-0.6779
	3	3	0.8164
	4	4	0.3617
	5	5	0.2527
	6	6	-0.0722

#### 4.2. Calibration Process

The CAPT 2012 tests were scaled and equated using the Rasch model. The WINSTEPS software was used to estimate the latent trait difficulty of each item on the test. WINSTEPS, written by Linacre (Mesa Press, 2005) was used to complete Rasch analyses. WINSTEPS is a WINDOWS-based program that is widely used for similar high stakes tests. WINSTEPS (based on the Rasch model), allows for the estimation of item difficulty for multiple-choice, open-ended, and extended response items on a single scale. Using these item difficulties, the model is able to estimate the ability (theta) of each student corresponding to each student's raw score.

All scaling and equating analyses were undertaken by three independent groups: Measurement Incorporated (MI), the contractor, the Connecticut State Department of Education (CSDE), and H. Jane Rogers and H. Swaminathan from the University of Connecticut (UCONN). Results were compared and cross-checked to the fourth decimal point to ensure accuracy.

The purpose of equating was to place the difficulty estimates of the items on the same scale as HS18 (CAPT 2011 Live). The equating was accomplished in the following steps:

1. For every content area, calibrate all items in 2012 OP (see Charts 1-4 for sample calibration data matrix). This step is a free run calibration. For RL, IW1, and IW2, 2 is subtracted from each score so that scores are on a scale from 0 to 10.

**Chart 1: Calibration Design for 2012 Mathematics**

Form HS19	HS19_MA1	HS19_MA2
-----------	----------	----------

**Note:**

HS19\_MA1 = Form HS19 Math Session 1

HS19\_MA2 = Form HS19 Math Session 2

**Chart 2: Calibration Design for 2012 Science**

Form HS19	HS19_SC1	HS19_SC2
-----------	----------	----------

**Note:**

HS19\_SC1 = Form HS19 Science Session 1

HS19\_SC2 = Form HS19 Science Session 2

**Chart 3: Calibration Design for 2012 Reading**

Form HS19	HS19_RI	HS19_RL
-----------	---------	---------

**Note:**

HS19\_RI = Form HS19 Reading for Information

HS19\_RL = Form HS19 Response to Literature

**Chart 4: Calibration Design for 2012 Writing**

HS19	HS19_ER	HS19_IW1	HS19_IW2
------	---------	----------	----------

**Note:**

HS19\_ER = Form HS19 Editing & Revising

HS19\_IW1 = Form HS19 Interdisciplinary Writing 1

HS19\_IW2 = Form HS19 Interdisciplinary Writing 2

2. Select the items linking HS19 (2012 live test) and HS18 (2011 live test). Do anchor evaluation using .3 rule between the estimates of difficulties from Step 1 and HS18 values (see Table 3 for the Rasch values of linking items between Form 19 and Form 18). This is an iterative process in which each item, starting with the one with the greatest absolute value difference, is removed until all items fulfill the criterion for inclusion. Using the remaining items the difference between the scale means from HS18 and Step 1 yields the equating constant. Table 4 shows the equating constants for Form 19 and Form 18.

**Table 4: 2012 CAPT Equating Constants**

Content Area	Equating Constant
Mathematics	-0.0894
Reading	0.1229
Science	0.0545
Writing	0.1843

3. Using the item output files from step 1 and anchoring their b-values, perform another run for each combination of forms, i.e., employ only those items from a given form in order to obtain theta

values for each group of students administered a particular form. For Reading and Writing, the appropriate weights were included in the second calibration (see Table 5).

**Table 5: Summary of Weighting for Reading and Writing**

Content/Subject	Unweighted Scale	% of Total Scale	Score Weight	Compute Formula	Weighted Scale
Reading for Information	0 - 24	50%	1.0		0 - 24
Response to Literature	2 - 12	50%	2.4	(RL - 2)*2.4	0 - 24
Total Reading	2 - 36				0 - 48
Editing & Revising	0 - 18	30%	1.0		0 - 18
Interdisciplinary Writing 1	2 - 12	35%	2.1	(IW1 - 2)*2.1	0 - 21
Interdisciplinary Writing 2	2 - 12	35%	2.1	(IW2 - 2)*2.1	0 - 21
Total Writing	4 - 42				0 - 60

4. Compute scale score (SS) and scale score standard error (SSE) for each form:

$$SS = \left( \frac{T + EQ - T_{mean}}{T_{SD}} \right) * 45 + 250 \text{ and } SSE = \frac{T_{err}}{T_{SD}} * 45$$

where

$T$  and  $T_{err}$  are the ability score and the standard error of the ability from the score file in Step 3 (for Reading and Writing) and Step 1 (for Mathematics and Science).

$EQ$  is the difference between the mean of difficulty estimates of the linking items on HS18 and mean of difficulty estimates of the linking items on HS19, called the equating constant. This value was obtained in Step 2.

$T_{mean}$  and  $T_{SD}$  are the scaling coefficients from base year of CAPT2 (see Table 6).

**Table 6: Scaling Coefficients from Base Year (CAPT2)**

Content Area	T_mean	T_SD
Mathematics	-0.2317	1.6051
Science	0.4077	0.9254
Reading	0.4843	1.2278
Writing	1.0931	1.1187

The minimum SS is set to 100 and the maximum SS is set to 400. Any SS less than 100 was reset to 100 and any SS greater than 400 was reset to 400.

Appendix B contains the results of raw scores, theta, and scale score for HS19. Please contact CSDE for other forms and combinations.

## Part 5: Test Statistics

### 5.1. Reliability

Reliability is a statistical index of the consistency of test performance over repeated trials. The simplest model for conveying the concept of reliability is to describe the test re-test method. If a test is administered to a group of examinees and then re-administered to the same examinees a short time later, the correlation of the scores across both test administrations estimates the reliability of the test. To measure reliability using a single administration, the test items are split using various techniques into half-length tests and those scores are then correlated. Cronbach's alpha estimates the lower-bound estimate of an infinite combination of split-halves and therefore is regarded as a very conservative method for assessing test reliability.

Table 7 summarizes reliability estimates for CAPT Mathematics, Science, Reading, and Writing. The reliability coefficients are based on Cronbach's alpha measure of internal consistency. When evaluating these results it is important to remember that reliability is partially a function of test length and thus reliability is likely to be greater for tests that have more items. Table 8 presents the mean and standard deviation of students' scale scores.

**Table 7: CAPT Cronbach's Alpha**

Form	Mathematics	Reading	Writing	Science
HS19	0.94	0.84	0.79	0.94

**Table 8: CAPT Scale Score Summary Statistics**

Subject	Mean	Standard Deviation
Mathematics	255.13	45.86
Reading	247.79	60.21
Writing	267.75	56.99
Science	260.09	50.79

### 5.2. Classification Consistency and Accuracy

Classification consistency (see Table 9) and accuracy (see Table 10) were measured using the IRT-Class program developed by [CASMA](#) (Center for Advanced Studies in Measurement and Assessment) at the University of Iowa. The decision consistency and accuracy was assessed based on the given ability distribution and the difficulty of the items (IRT parameters).

**Table 9: Classification Consistency**

Content Area	Overall Classification Consistency	Cut Below Basic - Basic	Cut Basic - Proficient	Cut Proficient - Goal	Cut Goal - Advanced
Mathematics	0.80449	0.95626	0.95588	0.94799	0.94108
Reading	0.90927	0.93335	0.96351	0.96324	0.96340
Science	0.79348	0.95933	0.95650	0.94350	0.92761
Writing	0.82889	0.88556	0.96736	0.96658	0.96663

**Table 10: Classification Accuracy**

<b>Content Area</b>	<b>Overall Classification Accuracy</b>	<b>Cut Below Basic - Basic</b>	<b>Cut Basic - Proficient</b>	<b>Cut Proficient - Goal</b>	<b>Cut Goal - Advanced</b>
Mathematics	0.85919	0.96936	0.96800	0.96340	0.95818
Reading	0.90727	0.95028	0.96334	0.97376	0.97210
Science	0.84902	0.97151	0.96861	0.96020	0.94794
Writing	0.86427	0.91650	0.97547	0.97932	0.96534

The results of the program show that for the most part, classifications are highly consistent (see Table 9). The consistency ratings at each cut score are generally in the upper 90s. The cumulative effect of applying all cut scores simultaneously yields an average consistency of around low 80s to low 90s. The classification accuracy examinations show that the accuracy indexes at each cut score are generally in the upper 90s (see Table 10).

The program also computes the false negative rates for the test, which in effect are an estimate of those students that may have been misclassified in a performance category lower than their true performance category. The results of the false negatives, found in Table 11, indicate that a very small number of students may have been negatively misclassified in this way. In contrast, the false positive rates, which are estimate of those students that may have been misclassified to a performance category higher than their true performance category, are presented in Table 12. The results indicate that a very small number of students may have been positively misclassified.

**Table 11: False Negative Classification**

<b>Content Area</b>	<b>Overall False Negative</b>	<b>Cut Below Basic - Basic</b>	<b>Cut Basic - Proficient</b>	<b>Cut Proficient - Goal</b>	<b>Cut Goal - Advanced</b>
Mathematics	0.07232	0.01415	0.01107	0.02152	0.02569
Reading	0.05051	0.02933	0.03346	0.00829	0.00593
Science	0.08211	0.01502	0.01199	0.02274	0.03277
Writing	0.07981	0.04077	0.01417	0.00983	0.03374

**Table 12: False Positive Classification**

<b>Content Area</b>	<b>Overall False Positive</b>	<b>Cut Below Basic - Basic</b>	<b>Cut Basic - Proficient</b>	<b>Cut Proficient - Goal</b>	<b>Cut Goal - Advanced</b>
Mathematics	0.06849	0.01649	0.02093	0.01508	0.01613
Reading	0.04222	0.02039	0.00320	0.01795	0.02197
Science	0.06887	0.01347	0.01940	0.01706	0.01929
Writing	0.05591	0.04273	0.01036	0.01085	0.00092

## Part 6: CAPT3 Standards

When standards were being established for first generation CAPT, a judgmental standard setting process called Modified Angoff (1971) was employed. Through that process, groups of educators who were familiar with the performance of students at a particular grade level in a particular content area were asked to predict how students who just meet a particular standard (e.g., goal standard) would perform on many different CAPT items. Using the judgment of these groups of educators in consideration with other validity checks, appropriate state goal and remedial standards were recommended by the Department and adopted by the State Board of Education. For the second generation CAPT (CAPT2), the standards were set using a method called Book Mark. In the procedure, all items in the test are arranged from easiest to most difficult. Then a group of educators are asked to mark up to the item at which a student at specific standard could respond to correctly. As in the first generation, the standards set by using the Book Mark method were adopted by the State Board of Education.

The third generation (CAPT3) standards were developed by carrying over the CAPT2 standards as well as department staff working with a CAPT3 Standards Advisory Panel composed of technical experts, district content experts and district research and testing specialists. The CAPT3 standards were set to be as rigorous as the CAPT2 standards. Transferring the standards allowed the Department to maintain the same performance standards for NCLB purposes. The purpose of this section is to summarize the procedures used to accomplish the task of carrying over the standards (see Cizek and Bunch, 2007, for a discussion of standard setting procedures).

In all content areas, the standards define the different academic performance levels. The state goal has been an important benchmark for judging the quality of education in Connecticut for more than a decade. The proficient standard is used for accountability purposes as required by No Child Left Behind (NCLB) to make determinations about Adequate Yearly Progress (AYP) and schools in need of improvement.

To continue to comply with the NCLB accountability requirements, the Connecticut State Department of Education (CSDE) carried over from the CAPT2 to the CAPT3 the following previously adopted achievement standards: Below Basic, Basic, Proficient, Goal and Advanced. The process of carrying over the standards was accomplished with an intergeneration linking study which included the equating of CAPT2 forms and CAPT3 forms. In addition to statistically linking the test generations, historical results from past CAPT2 administrations were taken into consideration as well as input from the CAPT Standards Review Panel composed of a diverse group of Connecticut educators, including curriculum directors, teachers and administrators.

The Standards Review Panel assisted in the identification of acceptable and valid test standards for each content area of CAPT3. The CAPT Standards Review Panel was given an overview of the CAPT3 including the content covered, score weighting, and reporting conventions. Differences between CAPT2 and CAPT3 were also discussed. Copies of the complete CAPT3 test booklets were available for reference. In addition, the procedures for carrying CAPT2 standards over to CAPT3 were presented in detail so that committee members would better understand their role in the process. They reviewed data from several related analyses and discussed implications from both an educational perspective and a technical perspective. They were asked particularly to provide input in the following three areas:

- Review the content of the CAPT, score weighting, and reporting conventions.
- Review results from the inter-generational linking procedure to ensure that standards are reasonable and appropriate across content areas; and
- Provide subjective input about the reasonableness and consistency of the standards for all content areas based on their content expertise and historical results from past test administrations.

All procedures were discussed with and approved by the Technical Advisory Committee (TAC) prior to implementation. The TAC is composed of nationally recognized experts in the measurement field. Finally, standards proposed by the standards review panel were presented to the State Board of Education for final approval. Standards were established based on scale scores (100-400) in four content areas: Mathematics, Science, Reading, and Writing.

Table 13 shows the range of scale scores in each performance category.

**Table 13: 2012 CAPT Achievement Levels and Scale Score Ranges**

Content Area	Scale Score Ranges				
	Below Basic	Basic	Proficient	Goal	Advanced
Mathematics	100 - 190	191 - 220	221 - 259	260 - 289	290 - 400
Science	100 - 189	190 - 214	215 - 264	265 - 294	295 - 400
Reading	100 - 173	174 - 204	205 - 250	251 - 282	283 - 400
Writing	100 - 181	182 - 209	210 - 249	250 - 285	286 - 400

## **Part 7: Validity**

According to the 1999 AERA, APA, NCME *Standards*, “It is helpful to consider the four phases leading from the original statement of purpose(s) to the final product: (a) delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured; (b) development and evaluation of the test specifications; (c) development, field testing, evaluation, and selection of the items and scoring guides and procedures; and (d) the assembly and evaluation of the test for operational use.

In the development and maintenance of CAPT each of these phases is carefully planned and implemented. The following sections detail the important psychometric procedures undertaken to ensure a strong validity argument for the use and interpretation of CAPT (Kane, 2006; Messick, 1989).

### **7.1. Content Validity Survey**

In order for the CAPT to serve its intended purposes, it is critical that users of the test results be confident that those results are meaningful. The test must measure those competencies that are critical to the decisions the test scores are informing.

A content validation study was conducted to examine the content validity of the CAPT for its intended applications. For this study, a survey of the strands proposed for the second generation CAPT was sent to approximately 4,000 Connecticut educators, parents, and other citizens. The purpose of the survey was to determine 1) the importance of the proposed Mathematics, Science, Reading Across the Disciplines, and Writing Across the Disciplines strands and 2) whether the strands are taught prior to the end of the 10<sup>th</sup> grade. The respondents characterized the strands as important educational outcomes to which students would be instructed prior to testing.

### **7.2. Scoring Quality Assurance Procedures Undertaken during Development**

Much of the following discussion applies to procedures undertaken during field testing and test construction phases of development work. Of course quality control is applied during the operational administration, but not with the aim of selecting or removing items.

In order to ensure the validity of inferences made from the CAPT tests there are quality control procedures in place for the scoring of the test. One such quality assurance component is to check the MC answer keys for MC items several times prior to test administration and one final time during the first run of live results. Items yielding low point biserial correlations are checked a final time for miskeying.

For constructed-response (CR) items, CAPT staff and contractor staff work with Connecticut educators to establish score boundaries in a process known as “range finding”. The score point examples and training sets so established are carried forward into operational scoring and elaborated with new samples of student responses. Reader training lasts up to several days, and readers must qualify by matching scores to several sets of prescored student responses. Once scoring begins, validity packets are used to maintain reader accuracy. These are packets of student responses with scores pre-assigned by CAPT staff and Connecticut educators. Readers periodically receive these packets, and their responses are compared to the pre-assigned scores. If a reader assigns too many discrepant scores, that reader is retrained or removed from the project. Other QA procedures include a 100% second read for the writing prompts (IW). There is a 20% second read for short answer and extended response items in mathematics and reading comprehension.

### **7.3. Item Quality Analysis Undertaken During Development**

Another part of assessing the quality and validity of inferences made from an instrument is to assess the quality of the items on the test. This quality is typically assessed by examining the classical item statistics as well as the potential for item bias. Item bias could lead to invalid inferences made for certain subgroups.

*Item specifications.* CAPT employs *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) as a primary source of guidance in the construction, field testing, and documentation of the tests. The introduction to the 1999 *Standards* best describes how those *Standards* are and will be used in the development and evaluation of CAPT tests:

Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. (*Standards*, p. 4)

Thus, the terms ‘target’ and ‘goal’ are used when referring to various psychometric properties of the tests. For example, while it is a goal of test development for each high school test to have a reliability coefficient of .90 or greater, it is not our intention to scrap a test with a reliability coefficient of .89. Instead, the test results would be published, along with the reliability coefficient and associated standard error of measurement.

*Item statistics.* Because the CAPT tests are used in making individual decisions about students, they must be very reliable, particularly at cut points (the score points that separate adjacent achievement categories). Target reliability coefficients of .90 (or higher) are therefore set for the important cut points of each test.

Other psychometric properties include item difficulty, item discrimination, and differential item functioning. General statistical targets are provided below:

*For Multiple-Choice (MC) Items*

Percent correct: greater than or equal to .25  
Point biserial correlation with total score: greater than or equal to .20  
Mantel-Haenszel: No Category C items (see below)

*For Constructed-Response (CR) Items*

Difficulty: any level as long as all score points are well represented  
Correlation with total score: greater than or equal to .20  
Generalized Mantel-Haenszel: No chi-square significant at .05 level of alpha

It should be pointed out that the point biserial correlations for MC items and the correlations for CR items refer to total scores of the field test form with the influence of the item in question removed.

*Differential item functioning.* The Mantel\_Haenszel statistic computes an odds ratio for each item that compares item performance for a reference group and a focal group (for whom bias may be an issue). Specifically, the M-H statistic is a ratio of the probability of success on an item for the reference group to the probability of success on the same item for the focal group. When the ratio is greater than one, the probability of success on the item favors the reference group over the focal group. Note that M-H and other methods for identifying statistical bias are flagging mechanisms that do not necessarily mean that the performance difference is due to unfairness in the item. Instead, the standard procedure is for the bias committee review the items to make a final judgmental determination as to whether or not the item is actually biased.

Since its introduction in the field of epidemiology in 1959, Mantel-Haenszel statistics have been employed by many test developers, and several refinements have been added. Educational Testing Service (ETS) uses the Mantel-Haenszel statistic and calculates a D statistic which permits grouping of test items into three categories (Zieky, 1993). The D statistic is a function of the case-control odds estimator of risk generated by SAS’s PROC FREQ. The D statistic is calculated as follows:

1.  $\alpha$  = case-control estimate of risk (odds ratio)
2.  $\beta$  = natural log of  $\alpha$
3.  $D = -2.35*\beta$

Camilli and Shepard (1994, p. 121) describe three categories of items with respect to D:

- A D does not significantly differ from zero using Mantel-Haenszel chi-square, or D's absolute value is less than 1
- B D significantly differs from 0 and D has either (a) an absolute value less than 1.5 or (b) an absolute value not significantly different from 1
- C D's absolute value is significantly greater than or equal to 1.5

Camilli and Shepard note that Category B items are typically investigated for potential bias, while Category C items are typically removed. Others treat Category C items only as candidates for elimination, pending a reprieve from the committee. In other words, Category C items are considered unusable unless specifically declared usable by the committee. It should be noted that an item that allowed a target group to break out of a pattern of trailing behind the reference group on all other items would tend to fall into Category C. The committee would likely want to keep such an item, in spite of its Mantel-Haenszel status.

DIF occurs when an item shows different results by group (e.g., by race, or sex) that cannot be explained by known differences in the overall achievement levels of the two groups. Overall achievement level is typically taken as scores on an operational test, assuming that the operational test is itself free of bias. While committee members are free to examine all field-tested items, they must review all items with a Category C rating. Unless the committee specifically calls for the inclusion of any such item, that item is removed from the pool.

#### **7.4. Equating Design**

A different CAPT form is used each year. In order to ensure that appropriate comparisons can be made from one form of the CAPT to another, test forms must be equivalent to each other. Care must be taken when test items are developed, when items are selected to create forms, when tests are administered, and when tests are scored to keep all conditions as similar as possible for one test form to another. Two important characteristics that must be similar across forms are the content that is measured and the difficulty of the test.

Part 4 of this report details the procedures used to equate and scale the CAPT tests. As mentioned above, three independent groups undertake the analyses and cross-check all analyses and results to ensure accuracy. Connecticut expends great effort and resources to maintain an assessment program that employs high quality psychometric standards and quality assurance.

## REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2<sup>nd</sup> ed., pp. 508-600), Washington, DC: American Council on Research.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4<sup>th</sup> ed., pp. 18-64). Westport, CT: American Council on Education/Praeger.
- Linacre, J. M., & Wright, B. D. (1993, 2006). *A user's guide to BIGSTEPS*. Chicago, IL: MESA Press.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3<sup>rd</sup> ed., pp. 13-103). New York: American Council on Education/Macmillan Publishing Company.
- Winsteps. (1991-2006©). Linacre, John M.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 337-364). Hillsdale, NJ: Lawrence Erlbaum Associates.

## Appendix A: Item Analysis

### Mathematics HS19 Item Analysis

#### Grid-in Items

PC = Proportion Correct

RPB = Point-Biserial correlation

#### Open-ended Items

Mean = Mean OE score

Corr = Item-total correlation

0 – 3 = Percent of students at each score point

Item	Type	Rasch	PC/Mean	RPB/Corr
1	OE	-0.9387	1.92	0.54
2	OE	0.3313	1.28	0.71
3	OE	0.4108	1.28	0.65
4	OE	1.3856	0.71	0.59
5	GR	-1.3981	0.73	0.31
6	GR	0.5252	0.42	0.56
7	GR	0.0346	0.50	0.57
8	GR	0.8525	0.36	0.59
9	GR	-0.8888	0.66	0.41
10	GR	0.2086	0.47	0.61
11	GR	0.3750	0.44	0.59
12	GR	-0.9726	0.67	0.55
13	GR	-0.5976	0.61	0.59
14	GR	0.8672	0.36	0.54
15	GR	0.9770	0.34	0.60
16	GR	0.3192	0.45	0.41
17	OE	0.2303	1.38	0.71
18	OE	-0.9035	1.98	0.59
19	OE	-0.0721	1.57	0.71
20	OE	0.7800	1.09	0.68
21	GR	-1.0837	0.69	0.30
22	GR	0.5289	0.42	0.60
23	GR	-0.8908	0.66	0.61
24	GR	-1.4694	0.74	0.47
25	GR	0.0479	0.50	0.53
26	GR	-0.5515	0.60	0.53
27	GR	0.0782	0.49	0.58
28	GR	0.0431	0.50	0.62

<b>Item</b>	<b>Type</b>	<b>Rasch</b>	<b>PC/Mean</b>	<b>RPB/Corr</b>
29	GR	-1.7917	0.78	0.38
30	GR	0.7628	0.38	0.49
31	GR	0.1675	0.48	0.56
32	GR	-0.2279	0.55	0.53

## Science HS19 Item Analysis

### Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

A – D = Proportion answering each distractor; answer key is shaded

### Open-ended Items

Mean = Mean OE score

Corr = Item-total correlation

0 – 3 = Percent of students at each score point

Item	Type	Rasch	PC/Mean	RPB/Corr
1	OE	0.1741	1.85	0.61
2	OE	-0.1531	2.05	0.58
3	OE	0.0513	1.90	0.67
4	MC	-0.5926	0.73	0.48
5	MC	-0.7285	0.75	0.40
6	MC	0.0258	0.62	0.48
7	MC	0.6184	0.50	0.39
8	MC	0.6167	0.50	0.27
9	MC	-1.0053	0.79	0.40
10	MC	-0.8189	0.77	0.37
11	MC	-0.5360	0.72	0.32
12	MC	0.0335	0.62	0.48
13	MC	0.5660	0.51	0.40
14	MC	0.2019	0.58	0.47
15	MC	0.1369	0.60	0.34
16	MC	-0.1696	0.66	0.30
17	MC	-1.1930	0.82	0.42
18	MC	0.0134	0.62	0.57
19	MC	0.2253	0.58	0.50
20	MC	0.4281	0.54	0.28
21	MC	0.9541	0.43	0.38
22	MC	0.7062	0.48	0.31
23	MC	0.5121	0.52	0.33
24	MC	-0.2966	0.68	0.56
25	MC	0.4801	0.53	0.29
26	MC	-0.9816	0.79	0.49
27	MC	0.5601	0.51	0.40
28	MC	-0.2293	0.67	0.49
29	MC	0.4843	0.53	0.34
30	MC	-0.3132	0.68	0.36

Item	Type	Rasch	PC/Mean	RPB/Corr
31	MC	-0.3361	0.69	0.54
32	OE	0.2306	1.74	0.54
33	OE	0.1202	1.83	0.56
34	MC	-0.8297	0.77	0.41
35	MC	0.4467	0.54	0.30
36	MC	-0.0701	0.64	0.28
37	MC	-0.6116	0.73	0.33
38	MC	-0.5153	0.72	0.49
39	MC	1.3714	0.35	0.37
40	MC	0.6390	0.50	0.33
41	MC	-1.0857	0.80	0.40
42	MC	1.1177	0.40	0.31
43	MC	-0.8033	0.76	0.37
44	MC	-0.3430	0.69	0.30
45	MC	1.3319	0.36	0.26
46	MC	-0.6005	0.73	0.33
47	MC	1.1545	0.39	0.27
48	MC	-0.5244	0.72	0.52
49	MC	0.6798	0.49	0.33
50	MC	0.9330	0.44	0.50
51	MC	0.2352	0.58	0.46
52	MC	0.8948	0.44	0.28
53	MC	0.5490	0.51	0.25
54	MC	0.1006	0.60	0.43
55	MC	-0.1162	0.65	0.51
56	MC	0.7425	0.48	0.42
57	MC	0.0489	0.61	0.38
58	MC	0.3787	0.55	0.44
59	MC	0.0441	0.61	0.46
60	MC	-0.0902	0.64	0.49
61	MC	0.7808	0.47	0.29
62	MC	-0.5920	0.73	0.52
63	MC	-0.2788	0.68	0.53
64	MC	-0.0934	0.64	0.48
65	MC	-1.1371	0.81	0.52

## Reading for Information HS19 Item Analysis

### Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

A – D = Proportion answering each distractor; answer key is shaded

### Open-ended Items

Mean = Mean OE score

Corr = Item-total correlation

0 – 2 = Percent of students at each score point

Item	Type	Rasch	PC/Mean	RPB/Corr
1	MC	0.1325	0.54	0.30
2	MC	-1.7697	0.84	0.33
3	MC	-1.0146	0.75	0.37
4	MC	0.0194	0.57	0.29
5	OE	0.8228	0.87	0.55
6	OE	0.8056	0.88	0.58
7	MC	-0.0117	0.57	0.25
8	MC	-0.7660	0.71	0.38
9	MC	0.0084	0.57	0.29
10	MC	-0.5299	0.67	0.39
11	OE	1.1393	0.77	0.53
12	OE	0.7960	0.87	0.54
13	MC	0.0051	0.57	0.32
14	MC	-0.5569	0.67	0.30
15	MC	-0.0969	0.59	0.43
16	MC	0.3962	0.49	0.38
17	OE	1.0461	0.79	0.59
18	OE	1.3095	0.71	0.56

### Editing and Revising HS19 Item Analysis

**Multiple-choice Items**

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

A – D = Proportion answering each distractor; answer key is shaded

Item	Type	Rasch	PC/Mean	RPB/Corr
1	MC	-0.7153	0.86	0.35
2	MC	-0.6726	0.85	0.23
3	MC	0.8129	0.63	0.24
4	MC	0.2680	0.72	0.40
5	MC	0.3301	0.71	0.26
6	MC	-0.1079	0.78	0.27
7	MC	-0.6188	0.85	0.30
8	MC	0.7205	0.65	0.24
9	MC	0.3812	0.71	0.39
10	MC	0.3792	0.71	0.34
11	MC	0.1072	0.75	0.36
12	MC	-0.6142	0.85	0.35
13	MC	0.5796	0.67	0.32
14	MC	-0.3554	0.82	0.40
15	MC	-0.5869	0.84	0.30
16	MC	-0.2594	0.80	0.34
17	MC	1.4319	0.51	0.25
18	MC	-0.2368	0.80	0.41

### Response to Literature and Interdisciplinary Writing HS19 Item Analysis

**Extended Response**

Mean = Mean EX score

Corr = Item-total correlation

2 – 12 = Percent of students at each point

	Type	Rasch	Mean	Corr	2	3	4	5	6	7	8	9	10	11	12
RL	EX	0.5998	7.05	0.63	0.01	0.01	0.06	0.08	0.23	0.19	0.27	0.11	0.05	0.01	0.00
IW1	EX	1.4312	7.67	0.72	0.02	0.01	0.04	0.05	0.11	0.13	0.31	0.17	0.12	0.03	0.01
IW2	EX	1.4115	7.61	0.73	0.02	0.02	0.04	0.05	0.12	0.15	0.27	0.18	0.11	0.03	0.01

**Appendix B: Raw, Theta, and Scale Scores**

**Raw, Theta, and Scale Scores for Mathematics HS19**

Raw Score	Theta	Scale Score
0	-5.3805	103
1	-4.1369	138
2	-3.3880	159
3	-2.9279	172
4	-2.5865	181
5	-2.3104	189
6	-2.0759	196
7	-1.8702	202
8	-1.6857	207
9	-1.5175	211
10	-1.3623	216
11	-1.2176	220
12	-1.0816	224
13	-0.9530	227
14	-0.8308	231
15	-0.7139	234
16	-0.6019	237

Raw Score	Theta	Scale Score
17	-0.4939	240
18	-0.3897	243
19	-0.2885	246
20	-0.1901	249
21	-0.0940	251
22	0.0001	254
23	0.0926	257
24	0.1838	259
25	0.2740	262
26	0.3637	264
27	0.4531	267
28	0.5426	269
29	0.6326	272
30	0.7233	274
31	0.8152	277
32	0.9087	279
33	1.0042	282

Raw Score	Theta	Scale Score
34	1.1023	285
35	1.2036	288
36	1.3088	291
37	1.4188	294
38	1.5347	297
39	1.6583	300
40	1.7916	304
41	1.9377	308
42	2.1013	313
43	2.2898	318
44	2.5155	325
45	2.8020	333
46	3.2025	344
47	3.8868	363
48	5.0825	396

**Raw, Theta, and Scale Scores for Science HS19**

Raw Score	Theta	Scale Score
0	-5.6162	100
1	-4.3991	100
2	-3.6871	100
3	-3.2629	100
4	-2.9564	100
5	-2.7144	101
6	-2.5132	111
7	-2.3401	119
8	-2.1875	126
9	-2.0507	133
10	-1.9263	139
11	-1.8118	145
12	-1.7056	150
13	-1.6063	155
14	-1.5129	159
15	-1.4244	164
16	-1.3404	168
17	-1.2601	172
18	-1.1832	175
19	-1.1092	179
20	-1.0378	182
21	-0.9687	186
22	-0.9018	189
23	-0.8366	192
24	-0.7731	195
25	-0.7112	198

Raw Score	Theta	Scale Score
26	-0.6504	201
27	-0.5909	204
28	-0.5323	207
29	-0.4747	210
30	-0.4179	212
31	-0.3617	215
32	-0.3060	218
33	-0.2508	221
34	-0.1960	223
35	-0.1414	226
36	-0.0870	229
37	-0.0327	231
38	0.0216	234
39	0.0760	237
40	0.1306	239
41	0.1856	242
42	0.2408	245
43	0.2965	247
44	0.3528	250
45	0.4098	253
46	0.4675	256
47	0.5261	258
48	0.5856	261
49	0.6463	264
50	0.7082	267
51	0.7715	270

Raw Score	Theta	Scale Score
52	0.8363	273
53	0.9028	277
54	0.9712	280
55	1.0417	283
56	1.1146	287
57	1.1901	291
58	1.2686	294
59	1.3505	298
60	1.4361	303
61	1.5261	307
62	1.6211	312
63	1.7219	317
64	1.8296	322
65	1.9456	327
66	2.0715	334
67	2.2097	340
68	2.3636	348
69	2.5381	356
70	2.7406	366
71	2.9838	378
72	3.2915	393
73	3.7169	400
74	4.4299	400
75	5.6477	400

**Raw, Theta, and Scale Scores for Reading HS19**

Raw Score	Theta	Scale Score
0	-5.5153	100
1	-4.4651	100
2	-3.8730	100
3	-3.5052	108
4	-3.2182	119
5	-2.9702	128
6	-2.7441	136
7	-2.5329	144
8	-2.3335	151
9	-2.1447	158
10	-1.9656	165
11	-1.7951	171
12	-1.6318	177
13	-1.4745	183
14	-1.3220	188
15	-1.1734	194
16	-1.0277	199

Raw Score	Theta	Scale Score
17	-0.8843	204
18	-0.7430	210
19	-0.6030	215
20	-0.4641	220
21	-0.3258	225
22	-0.1874	230
23	-0.0485	235
24	0.0918	240
25	0.2343	245
26	0.3799	251
27	0.5294	256
28	0.6837	262
29	0.8437	268
30	1.0101	274
31	1.1830	280
32	1.3626	287
33	1.5488	294

Raw Score	Theta	Scale Score
34	1.7414	301
35	1.9406	308
36	2.1470	315
37	2.3625	323
38	2.5897	332
39	2.8326	341
40	3.0962	350
41	3.3859	361
42	3.7064	373
43	4.0598	386
44	4.4475	400
45	4.8806	400
46	5.4023	400
47	6.1750	400
48	7.4063	400

**Raw, Theta, and Scale Scores for Writing HS19**

Raw Score	Theta	Scale Score
0	-4.5922	100
1	-3.3894	100
2	-2.7015	105
3	-2.3040	121
4	-2.0256	132
5	-1.8120	141
6	-1.6387	147
7	-1.4926	153
8	-1.3658	158
9	-1.2532	163
10	-1.1513	167
11	-1.0577	171
12	-0.9705	174
13	-0.8883	178
14	-0.8102	181
15	-0.7351	184
16	-0.6625	187
17	-0.5915	190
18	-0.5220	192
19	-0.4532	195
20	-0.3849	198

Raw Score	Theta	Scale Score
21	-0.3167	201
22	-0.2482	203
23	-0.1791	206
24	-0.1092	209
25	-0.0381	212
26	0.0345	215
27	0.1089	218
28	0.1854	221
29	0.2644	224
30	0.3462	227
31	0.4313	231
32	0.5200	234
33	0.6130	238
34	0.7106	242
35	0.8135	246
36	0.9223	250
37	1.0377	255
38	1.1601	260
39	1.2899	265
40	1.4277	271
41	1.5733	277

Raw Score	Theta	Scale Score
42	1.7266	283
43	1.8870	289
44	2.0542	296
45	2.2278	303
46	2.4079	310
47	2.5948	318
48	2.7896	326
49	2.9933	334
50	3.2074	342
51	3.4325	351
52	3.6694	361
53	3.9191	371
54	4.1836	382
55	4.4684	393
56	4.7849	400
57	5.1558	400
58	5.6344	400
59	6.3899	400
60	7.6306	400