

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

**Prepared by  
Irene Hendrawan & Arianto Wibowo**

**September 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. 2013 CAPT Test Design.....	1
1.3. 2013 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. 2013 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 2013 Mathematics</b> .....	6
<b>Chart 2: Calibration Design for 2013 Science</b> .....	7
<b>Chart 3: Calibration Design for 2013 Reading</b> .....	7
<b>Chart 4: Calibration Design for 2013 Writing</b> .....	7

## List of Tables

<b>Table 1: 2013 CAPT Operational Test Design</b> .....	1
<b>Table 2: Summary of Item Analysis Form HS20</b> .....	4
<b>Table 3: 2013 Embedded Linking Items</b> .....	5
<b>Table 4: 2013 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: 2013 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 2013 administration was the seventh operational (OP) administration of CAPT3.

## 1.2. 2013 CAPT Test Design

The spring 2013 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: 2013 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. 2013 CAPT Test Forms**

In the 2013 administration, two main forms were available for administration: Form HS20, which is the live form taken by most of the students, and Form HS0, which was available for breach situations. Moreover, Form HS0 will be used as a breach form in subsequent years of 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 forms in 2013. The pilot forms were administered to schools stratified by last year's achievement scores. CSDE's rationale for stratifying the pilot 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 (HS20 or HS0) was given the corresponding test session or subtest (HS20 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 content assessment 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 earlier generations of 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 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 result of item analysis, , 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 HS20**

Subject	Rasch		P-value		Point Biserial	
	Mean	Std	Mean	Std	Mean	Std
Mathematics	-0.10	1.02	0.75	0.48	0.54	0.09
Reading for Information	0.21	0.99	0.67	0.13	0.41	0.11
Response to Literature	0.84	-	6.85	-	0.62	-
Editing and Revising	0.01	0.66	0.75	0.10	0.33	0.08
Interdisciplinary Writing	1.38	0.05	7.75	0.08	0.73	0.01
Science	0.11	0.52	0.71	0.36	0.41	0.10

## Part 4: Scaling and Equating

### 4.1 2013 CAPT Linking Items

The 2013 CAPT Mathematics, Science, Reading for Information, and Editing & Revising tests were equated with the 2012 CAPT (HS19) subtests by embedding linking items. Linking items were counted toward students' scores.

The Live form of the 2013 CAPT (HS20) 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 2013 and 2012 tests.

**Table 3: 2013 Embedded Linking Items**

Content Area	Form HS20 Item Position	Form HS19 Item Position	Rasch Form HS19
Mathematics	9	9	-0.8888
	10	10	0.2086
	11	11	0.375
	12	12	-0.9726
	13	13	-0.5976
	14	14	0.8672
	21	21	-1.0837
	22	22	0.5289
	24	24	-1.4694
	27	27	0.0782
	29	29	-1.7917
	31	31	0.1675
Science	4	4	-0.5981
	18	18	0.0074
	19	19	0.2247
	23	23	0.5126
	24	24	-0.3003
	27	27	0.5649
	41	41	-1.0913
	42	42	1.1216
	56	56	0.7438
	57	57	0.0508
58	58	0.3808	

Content Area	Form HS20 Item Position	Form HS19 Item Position	Rasch Form HS19
	60	60	-0.0901
	61	61	0.7853
	62	62	-0.5953
	65	65	-1.1462
Reading	1	1	0.1327
	2	2	-1.7714
	3	3	-1.0165
	4	4	0.0211
	5	5	0.8227
	6	6	0.805
Writing	13	13	0.5784
	14	14	-0.3639
	15	15	-0.5828
	16	16	-0.259
	17	17	1.4347
	18	18	-0.2428

#### 4.2. Calibration Process

The CAPT 2013 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 HS19 (CAPT 2012 Live). The equating was accomplished in the following steps:

1. For every content area, calibrate all items in 2013 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 2013 Mathematics**

Form HS20	HS20_MA1	HS20_MA2
-----------	----------	----------

**Note:**

HS20\_MA1 = Form HS20 Math Session 1

HS20\_MA2 = Form HS20 Math Session 2

### Chart 2: Calibration Design for 2013 Science

Form HS20	HS20_SC1	HS20_SC2
-----------	----------	----------

**Note:**

HS20\_SC1 = Form HS20 Science Session 1

HS20\_SC2 = Form HS20 Science Session 2

### Chart 3: Calibration Design for 2013 Reading

Form HS20	HS20_RI	HS20_RL
-----------	---------	---------

**Note:**

HS20\_RI = Form HS20 Reading for Information

HS20\_RL = Form HS20 Response to Literature

### Chart 4: Calibration Design for 2013 Writing

HS20	HS20_ER	HS20_IW1	HS20_IW2
------	---------	----------	----------

**Note:**

HS20\_ER = Form HS20 Editing & Revising

HS20\_IW1 = Form HS20 Interdisciplinary Writing 1

HS20\_IW2 = Form HS20 Interdisciplinary Writing 2

2. Select the items linking HS20 (2013 live test) and HS19 (2012 live test). Do anchor evaluation using .3 rule between the estimates of difficulties from Step 1 and HS19 values (see Table 3 for the Rasch values of linking items between Form HS20 and Form HS19). 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 HS19 and Step 1 yields the equating constant. Table 4 shows the equating constants for Form HS20 and Form HS19.

**Table 4: 2013 CAPT Equating Constants**

Content Area	Equating Constant
Mathematics	-0.0981
Reading	0.2458
Science	0.1072
Writing	0.1437

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 HS19 and mean of difficulty estimates of the linking items on HS20, 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 HS20. 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
HS20	0.94	0.84	0.80	0.94

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

Subject	Mean	Standard Deviation
Mathematics	254.59	45.57
Reading	246.54	46.48
Writing	266.69	50.32
Science	263.07	51.07

### 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	0.80181	0.96332	0.95446	0.94759	0.93442
Reading	0.87904	0.90877	0.96768	0.96840	0.96842
Science	0.79597	0.95948	0.95513	0.94153	0.93360
Writing	0.85123	0.90914	0.96474	0.96180	0.96357

**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	0.85889	0.97397	0.96755	0.96315	0.95403
Reading	0.88636	0.93369	0.96139	0.97842	0.97842
Science	0.85117	0.97141	0.96764	0.95824	0.95333
Writing	0.87120	0.92837	0.96434	0.97405	0.96593

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 are estimates of the proportion of 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 portion of students may have been negatively misclassified in this way. In contrast, the false positive rates, which are estimates of the proportion of 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	0.06296	0.00998	0.01293	0.01663	0.02347
Reading	0.06753	0.03549	0.03653	0.00778	0.00679
Science	0.07374	0.01399	0.01227	0.02569	0.02205
Writing	0.08360	0.04119	0.02721	0.01220	0.03036

**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	0.07815	0.01605	0.01952	0.02022	0.02250
Reading	0.04611	0.03081	0.00208	0.01380	0.01479
Science	0.07509	0.01461	0.02009	0.01607	0.02462
Writing	0.04320	0.02844	0.00645	0.01175	0.00170

## 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 area; 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: 2013 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 focus 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 HS20 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	-0.2000	1.55	0.63
2	OE	-1.5210	2.18	0.54
3	OE	1.1200	0.85	0.62
4	OE	1.3388	0.78	0.60
5	GR	-0.4798	0.59	0.56
6	GR	-1.4116	0.73	0.47
7	GR	-0.8032	0.64	0.39
8	GR	1.7066	0.23	0.32
9	GR	-0.9426	0.66	0.45
10	GR	0.6268	0.40	0.56
11	GR	0.3281	0.45	0.59
12	GR	-0.9377	0.66	0.53
13	GR	-0.4523	0.59	0.61
14	GR	0.9407	0.35	0.53
15	GR	1.7509	0.22	0.52
16	GR	1.2113	0.30	0.53
17	OE	-1.1917	2.18	0.64
18	OE	-0.4102	1.57	0.64
19	OE	0.9073	0.97	0.61
20	OE	1.0913	0.97	0.65
21	GR	-1.1943	0.70	0.31
22	GR	0.6028	0.40	0.60
23	GR	-1.0494	0.68	0.57
24	GR	-1.4181	0.73	0.49
25	GR	-1.1233	0.69	0.55
26	GR	-0.3610	0.57	0.60
27	GR	-0.3375	0.57	0.63
28	GR	0.4237	0.43	0.45

<b>Item</b>	<b>Type</b>	<b>Rasch</b>	<b>PC/Mean</b>	<b>RPB/Corr</b>
29	GR	-1.7753	0.78	0.39
30	GR	0.5320	0.41	0.60
31	GR	-0.0162	0.51	0.56
32	GR	-0.0945	0.52	0.65

## Science HS20 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	-0.5413	2.20	0.57
2	OE	-0.0692	1.99	0.55
3	OE	0.2918	1.69	0.51
4	MC	-0.5634	0.73	0.51
5	MC	-0.4155	0.71	0.36
6	MC	0.7424	0.49	0.40
7	MC	-0.3472	0.70	0.42
8	MC	-0.5169	0.73	0.36
9	MC	-0.8338	0.78	0.40
10	MC	0.0977	0.62	0.39
11	MC	0.7928	0.48	0.20
12	MC	1.2235	0.39	0.35
13	MC	0.3249	0.57	0.38
14	MC	0.3223	0.57	0.16
15	MC	0.6302	0.51	0.34
16	MC	0.0335	0.63	0.30
17	MC	0.1243	0.61	0.51
18	MC	-0.0360	0.64	0.56
19	MC	0.0555	0.62	0.54
20	MC	0.2154	0.59	0.29
21	MC	0.5445	0.53	0.35
22	MC	-0.2037	0.67	0.37
23	MC	0.3152	0.57	0.36
24	MC	-0.3109	0.69	0.55
25	MC	0.4569	0.55	0.33
26	MC	-0.1139	0.66	0.46
27	MC	0.6294	0.51	0.40
28	MC	0.5941	0.52	0.37
29	MC	-0.3827	0.70	0.50
30	MC	0.1647	0.60	0.53

Item	Type	Rasch	PC/Mean	RPB/Corr
31	MC	-0.0775	0.65	0.48
32	OE	0.0533	1.91	0.56
33	OE	0.6035	1.67	0.58
34	MC	0.3369	0.57	0.27
35	MC	0.5295	0.53	0.43
36	MC	0.2629	0.58	0.16
37	MC	0.4310	0.55	0.37
38	MC	0.1406	0.61	0.40
39	MC	-0.2681	0.68	0.43
40	MC	-0.4332	0.71	0.33
41	MC	-0.9927	0.80	0.39
42	MC	1.1340	0.41	0.33
43	MC	0.2284	0.59	0.37
44	MC	0.0921	0.62	0.31
45	MC	-0.8729	0.78	0.37
46	MC	0.3719	0.56	0.48
47	MC	0.7765	0.48	0.50
48	MC	0.1631	0.60	0.44
49	MC	0.9249	0.45	0.45
50	MC	0.7951	0.48	0.33
51	MC	0.1307	0.61	0.31
52	MC	-0.1626	0.67	0.45
53	MC	-0.0561	0.65	0.44
54	MC	-0.3626	0.70	0.50
55	MC	-0.4598	0.72	0.48
56	MC	0.9926	0.44	0.38
57	MC	0.1466	0.61	0.36
58	MC	0.2757	0.58	0.44
59	MC	-0.7007	0.76	0.45
60	MC	-0.1394	0.66	0.47
61	MC	0.7718	0.48	0.28
62	MC	-0.4556	0.72	0.49
63	MC	0.5665	0.52	0.51
64	MC	0.2493	0.59	0.43
65	MC	-1.2525	0.83	0.50

## Reading for Information HS20 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	0.1372	0.55	0.32
2	MC	-1.7028	0.84	0.33
3	MC	-0.9915	0.75	0.37
4	MC	0.0562	0.56	0.29
5	OE	0.8819	0.87	0.52
6	OE	0.6124	0.95	0.52
7	MC	-0.3660	0.64	0.21
8	MC	-0.5342	0.67	0.42
9	MC	-0.0500	0.58	0.39
10	MC	-0.4708	0.66	0.24
11	OE	1.4848	0.76	0.47
12	OE	1.3869	0.71	0.53
13	MC	-0.2583	0.62	0.46
14	MC	0.9937	0.39	0.37
15	MC	-0.8572	0.72	0.44
16	MC	0.0907	0.56	0.40
17	OE	1.7225	0.64	0.58
18	OE	1.6979	0.57	0.57

### Editing and Revising HS20 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	0.0046	0.76	0.46
2	MC	0.3045	0.72	0.33
3	MC	-0.3898	0.82	0.25
4	MC	0.6026	0.67	0.17
5	MC	-0.2925	0.81	0.42
6	MC	-0.7423	0.86	0.39
7	MC	0.8638	0.62	0.29
8	MC	0.3485	0.71	0.23
9	MC	0.4050	0.70	0.28
10	MC	-0.5818	0.84	0.45
11	MC	-1.3348	0.91	0.41
12	MC	0.3659	0.71	0.30
13	MC	0.4073	0.70	0.32
14	MC	-0.2201	0.80	0.37
15	MC	-0.5682	0.84	0.28
16	MC	-0.3192	0.81	0.35
17	MC	1.4565	0.50	0.24
18	MC	-0.1914	0.79	0.42

### Response to Literature and Interdisciplinary Writing HS20 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	0.8368	6.85	0.62	0.02	0.01	0.08	0.09	0.23	0.18	0.27	0.09	0.04	0.01	0.00
IW1	EX	1.3434	7.80	0.72	0.02	0.01	0.04	0.04	0.10	0.12	0.31	0.17	0.14	0.04	0.01
IW2	EX	1.4121	7.69	0.73	0.02	0.01	0.04	0.04	0.11	0.14	0.32	0.18	0.11	0.02	0.01

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

**Raw, Theta, and Scale Scores for Mathematics HS20**

Raw Score	Theta	Scale Score
0	-5.7397	100
1	-4.4705	128
2	-3.6890	150
3	-3.2013	164
4	-2.8373	174
5	-2.5430	182
6	-2.2940	189
7	-2.0767	196
8	-1.8827	201
9	-1.7062	206
10	-1.5433	210
11	-1.3908	215
12	-1.2465	219
13	-1.1086	223
14	-0.9757	226
15	-0.8469	230
16	-0.7214	234

Raw Score	Theta	Scale Score
17	-0.5986	237
18	-0.4781	240
19	-0.3596	244
20	-0.2430	247
21	-0.1281	250
22	-0.0150	253
23	0.0964	256
24	0.2062	260
25	0.3144	263
26	0.4210	266
27	0.5264	269
28	0.6304	271
29	0.7336	274
30	0.8361	277
31	0.9385	280
32	1.0412	283
33	1.1447	286

Raw Score	Theta	Scale Score
34	1.2497	289
35	1.3571	292
36	1.4678	295
37	1.5832	298
38	1.7045	302
39	1.8339	305
40	1.9736	309
41	2.1273	313
42	2.2997	318
43	2.4986	324
44	2.7369	330
45	3.0385	339
46	3.4573	351
47	4.1641	370
48	5.3787	400

**Raw, Theta, and Scale Scores for Science HS20**

Raw Score	Theta	Scale Score
0	-5.6417	100
1	-4.4253	100
2	-3.7143	100
3	-3.2910	100
4	-2.9856	100
5	-2.7446	102
6	-2.5445	112
7	-2.3725	120
8	-2.2211	127
9	-2.0853	134
10	-1.9619	140
11	-1.8484	146
12	-1.7431	151
13	-1.6446	155
14	-1.5518	160
15	-1.4640	164
16	-1.3804	168
17	-1.3006	172
18	-1.2239	176
19	-1.1500	179
20	-1.0787	183
21	-1.0095	186
22	-0.9423	190
23	-0.8768	193
24	-0.8128	196
25	-0.7501	199

Raw Score	Theta	Scale Score
26	-0.6886	202
27	-0.6282	205
28	-0.5685	208
29	-0.5098	211
30	-0.4517	213
31	-0.3940	216
32	-0.3369	219
33	-0.2802	222
34	-0.2237	225
35	-0.1674	227
36	-0.1112	230
37	-0.0550	233
38	0.0012	235
39	0.0576	238
40	0.1142	241
41	0.1711	244
42	0.2284	246
43	0.2861	249
44	0.3444	252
45	0.4033	255
46	0.4630	258
47	0.5235	261
48	0.5850	264
49	0.6475	267
50	0.7113	270
51	0.7763	273

Raw Score	Theta	Scale Score
52	0.8429	276
53	0.9111	280
54	0.9812	283
55	1.0534	287
56	1.1279	290
57	1.2051	294
58	1.2852	298
59	1.3687	302
60	1.4561	306
61	1.5478	311
62	1.6447	315
63	1.7475	320
64	1.8573	326
65	1.9755	331
66	2.1039	338
67	2.2449	345
68	2.4020	352
69	2.5800	361
70	2.7867	371
71	3.0348	383
72	3.3482	398
73	3.7803	400
74	4.5013	400
75	5.7256	400

**Raw, Theta, and Scale Scores for Reading HS20**

Raw Score	Theta	Scale Score
0	-5.3174	100
1	-4.2189	100
2	-3.6135	109
3	-3.2542	122
4	-2.9861	132
5	-2.7631	140
6	-2.5658	147
7	-2.3844	154
8	-2.2133	160
9	-2.0498	166
10	-1.8920	172
11	-1.7389	178
12	-1.5896	183
13	-1.4435	188
14	-1.3000	194
15	-1.1584	199
16	-1.0182	204

Raw Score	Theta	Scale Score
17	-0.8790	209
18	-0.7402	214
19	-0.6016	219
20	-0.4626	224
21	-0.3230	229
22	-0.1821	235
23	-0.0393	240
24	0.1062	245
25	0.2553	251
26	0.4091	256
27	0.5686	262
28	0.7349	268
29	0.9086	275
30	1.0903	281
31	1.2797	288
32	1.4758	295
33	1.6774	303

Raw Score	Theta	Scale Score
34	1.8834	310
35	2.0931	318
36	2.3069	326
37	2.5265	334
38	2.7545	342
39	2.9940	351
40	3.2485	360
41	3.5206	370
42	3.8115	381
43	4.1224	392
44	4.4582	400
45	4.8354	400
46	5.3006	400
47	6.0136	400
48	7.1946	400

**Raw, Theta, and Scale Scores for Writing HS20**

Raw Score	Theta	Scale Score
0	-4.6491	100
1	-3.4457	100
2	-2.7578	101
3	-2.3610	117
4	-2.0834	128
5	-1.8706	137
6	-1.6981	144
7	-1.5526	149
8	-1.4263	154
9	-1.3141	159
10	-1.2126	163
11	-1.1192	167
12	-1.0322	170
13	-0.9501	174
14	-0.8719	177
15	-0.7968	180
16	-0.7240	183
17	-0.6529	186
18	-0.5830	188
19	-0.5139	191
20	-0.4451	194

Raw Score	Theta	Scale Score
21	-0.3763	197
22	-0.3071	199
23	-0.2372	202
24	-0.1664	205
25	-0.0942	208
26	-0.0204	211
27	0.0554	214
28	0.1334	217
29	0.2140	220
30	0.2976	224
31	0.3845	227
32	0.4753	231
33	0.5704	235
34	0.6704	239
35	0.7759	243
36	0.8874	248
37	1.0058	252
38	1.1316	257
39	1.2652	263
40	1.4069	268
41	1.5568	274

Raw Score	Theta	Scale Score
42	1.7146	281
43	1.8797	287
44	2.0518	294
45	2.2308	302
46	2.4171	309
47	2.6122	317
48	2.8179	325
49	3.0366	334
50	3.2710	343
51	3.5226	354
52	3.7914	364
53	4.0758	376
54	4.3742	388
55	4.6890	400
56	5.0297	400
57	5.4186	400
58	5.9087	400
59	6.6694	400
60	7.9097	400