

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

**Prepared by  
Irene Hendrawan & Arianto Wibowo**

**November 2009**



# Table of Contents

<b>Table of Contents</b> .....	2
<b>List of Charts</b> .....	3
<b>List of Tables</b> .....	4
<b>Part 1: Introduction</b> .....	5
1.1. General description of CAPT.....	5
1.2. 2009 CAPT Test Design.....	5
1.3. 2009 CAPT Test Forms.....	6
<b>Part 2: Test Development</b> .....	7
<b>Part 3: Item Level Statistics</b> .....	8
<b>Part 4: Scaling and Equating</b> .....	9
4.1. 2009 CAPT Linking Items.....	9
4.2. Calibration Process.....	10
<b>Part 5: Test Statistics</b> .....	13
5.1. Reliability.....	13
5.2. Classification Consistency and Accuracy.....	13
<b>Part 6: CAPT3 Standards</b> .....	15
<b>Part 7: Validity</b> .....	17
7.1. Content Validity Survey.....	17
7.2. Scoring Quality Assurance Procedures Undertaken during Development.....	17
7.3. Item Quality Analysis Undertaken During Development.....	17
7.4. Equating Design.....	19
<b>References</b> .....	20
<b>Appendix A: Item Analysis</b> .....	21
<b>Appendix B: Raw, Theta, and Scale Scores</b> .....	27

## List of Charts

<b>Chart 1: Calibration Design for 2009 Mathematics</b> .....	10
<b>Chart 2: Calibration Design for 2009 Science</b> .....	11
<b>Chart 3: Calibration Design for 2009 Reading</b> .....	11
<b>Chart 4: Calibration Design for 2009 Writing</b> .....	11

## List of Tables

<b>Table 1: 2009 CAPT Operational Test Design</b> .....	5
<b>Table 2: Summary of Item Analysis Form HS16</b> .....	8
<b>Table 3: 2009 Embedded Linking Items</b> .....	9
<b>Table 4: 2009 CAPT Equating Constants</b> .....	11
<b>Table 5: Summary of Weighting for Reading and Writing</b> .....	11
<b>Table 6: Scaling Coefficients from Base Year (CAPT2)</b> .....	12
<b>Table 7: CAPT Cronbach’s Alpha</b> .....	13
<b>Table 8: CAPT Scale Score Summary Statistics</b> .....	13
<b>Table 9: Classification Consistency</b> .....	13
<b>Table 10: Classification Accuracy</b> .....	14
<b>Table 11: False Negative Classification</b> .....	14
<b>Table 12: False Positive Classification</b> .....	14
<b>Table 13: 2009 CAPT Achievement Levels and Scale Score Ranges</b> .....	16

# 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 2009 administration was the third operational (OP) administration of CAPT3.

## 1.2. 2009 CAPT Test Design

The spring 2009 administration comprises the following content areas:

1. Mathematics  
Mathematics (MA) has thirty-two operational -- 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 an EX item with a 2-12 score scale (sum of two rater scores on a 1-6 scale).

**Table 1: 2009 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. 2009 CAPT Test Forms**

In the 2009 administration, two main forms were available for administration: Form HS16, 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.

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 (HS16 or HS0) was given the corresponding test session or subtest (HS16 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 item analysis (item quality) data 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.

**Item Discriminations:** The point biserial correlation or item-total correlations measure the strength of the relationship between the particular item score and the total 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. The percent of students at each score point is presented for extended response (2-12).

**Table 2: Summary of Item Analysis Form HS16**

Subject	Rasch		P-value		Point Biserial	
	Mean	Std	Mean	Std	Mean	Std
Mathematics	0.0118	0.8617	0.71	0.43	0.57	0.09
Reading for Information	-0.1661	1.1713	0.73	0.20	0.40	0.10
Response to Literature	0.4270		7.11		0.62	
Editing and Revising	-0.0259	1.1011	0.72	0.17	0.30	0.07
Interdisciplinary Writing	1.0243	0.0577	7.87	0.03	0.72	0.01
Science	-0.0390	0.6113	0.68	0.33	0.42	0.08

## Part 4: Scaling and Equating

### 4.1 2009 CAPT Linking Items

The 2009 CAPT Mathematics, Science, Reading for Information, and Editing & Revising tests were equated with the 2008 CAPT (HS15) subtests by embedding linking items. Linking items were counted toward students' scores.

The Live form of the 2009 CAPT (HS16) 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 2009 and 2008 tests.

**Table 3: 2009 Embedded Linking Items**

Content Area	Form HS16 Item Position	Form HS15 Item Position	Item Type	Rasch Form HS15
Mathematics	8	8	GR	0.7846
	9	9	GR	-0.1127
	10	10	GR	0.6468
	11	11	GR	1.2399
	13	13	GR	0.6796
	15	15	GR	0.4031
	21	21	GR	-2.5973
	22	22	GR	0.4235
	24	24	GR	0.1049
	27	27	GR	-0.1490
	28	28	GR	0.1403
	31	31	GR	0.1919
Science	17	17	MC	-0.3750
	20	20	MC	0.2835
	24	24	MC	-0.0138
	25	25	MC	-0.5401
	27	27	MC	0.0292
	28	29	MC	0.2054
	29	30	MC	0.4534
	39	9	MC	-0.1538
	40	40	MC	-0.6061
	41	41	MC	-0.2225
	42	42	MC	0.2121
	55	55	MC	-1.0039
	57	57	MC	0.1338

Content Area	Form HS16 Item Position	Form HS15 Item Position	Item Type	Rasch Form HS15
	58	58	MC	0.6887
	59	59	MC	0.0640
Reading	13	13	MC	-1.4871
	14	14	MC	0.7592
	15	15	MC	-1.0903
	16	16	MC	-0.6005
	17	17	OE	1.5988
	18	18	OE	2.0515
Writing	7	7	MC	-1.7100
	8	8	MC	1.4166
	9	9	MC	1.3188
	10	10	MC	0.4547
	11	11	MC	-0.9919
	12	12	MC	1.4881

#### 4.2. Calibration Process

The CAPT 2009 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 (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 HS15 (CAPT 2008 Live). The equating was accomplished in the following steps:

1. For every content area, concurrently calibrate the 2009 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 2009 Mathematics**

Form HS16	HS16_MA1	HS16_MA2
-----------	----------	----------

**Note:**

HS16\_MA1 = Form HS16 Math Session 1

HS16\_MA2 = Form HS16 Math Session 2

**Chart 2: Calibration Design for 2009 Science**

Form HS16	HS16_SC1	HS16_SC2
-----------	----------	----------

**Note:**

HS16\_SC1 = Form HS16 Science Session 1

HS16\_SC2 = Form HS16 Science Session 2

**Chart 3: Calibration Design for 2009 Reading**

Form HS16	HS16_RI	HS16_RL
-----------	---------	---------

**Note:**

HS16\_RI = Form HS16 Reading for Information

HS16\_RL = Form HS16 Response to Literature

**Chart 4: Calibration Design for 2009 Writing**

HS16	HS16_ER	HS16_IW1	HS16_IW2
------	---------	----------	----------

**Note:**

HS16\_ER = Form HS16 Editing & Revising

HS16\_IW1 = Form HS16 Interdisciplinary Writing 1

HS16\_IW2 = Form HS16 Interdisciplinary Writing 2

2. Select the items linking HS16 (2009 live test) and HS15 (2008 live test). Do anchor evaluation using .3 rule between the estimates of difficulties from Step 1 and HS15 values (see Table 3 for the Rasch values of linking items). 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 HS15 and Step 1 yields the equating constant. Table 4 shows the equating constants.

**Table 4: 2009 CAPT Equating Constants**

Content Area	Equating Constant
Mathematics	0.0118
Reading	-0.1349
Science	-0.0390
Writing	0.0791

3. Using the item output files from step 1 and anchoring these 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 (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

Content/Subject	Unweighted Scale	% of Total Scale	Score Weight	Compute Formula	Weighted Scale
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 forms:

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

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 HS15 and mean of difficulty estimates of the linking items on HS16, 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 HS16. 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 clusters 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
HS16	0.945	0.832	0.777	0.937

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

Subject	Mean	Standard Deviation
Mathematics	251.53	47.45
Reading	245.10	46.88
Writing	258.48	47.32
Science	254.62	50.41

### 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 1	Cut 2	Cut 3	Cut 4
Mathematics	0.79696	0.95783	0.94308	0.94972	0.94439
Reading	0.92907	0.94493	0.96363	0.96355	0.96357
Science	0.78873	0.95587	0.95144	0.94146	0.93291
Writing	0.92689	0.96143	0.96344	0.96254	0.96265

**Table 10: Classification Accuracy**

<b>Content Area</b>	<b>Overall Classification Accuracy</b>	<b>Cut 1</b>	<b>Cut 2</b>	<b>Cut 3</b>	<b>Cut 4</b>
Mathematics	0.85094	0.96873	0.95747	0.96409	0.96044
Reading	0.92478	0.95263	0.97077	0.97423	0.97397
Science	0.84557	0.96906	0.96512	0.95875	0.95183
Writing	0.94325	0.97099	0.97382	0.97342	0.97278

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. This tends to tail off at the highest cut score (i.e., the upper end of the distributions). 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 (see Table 10), similarly, that the accuracy ratings at each cut score are generally in the upper 90s.

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. Table 12 shows the false positive classification.

**Table 11: False Negative Classification**

<b>Content Area</b>	<b>Overall False Negative</b>	<b>Cut 1</b>	<b>Cut 2</b>	<b>Cut 3</b>	<b>Cut 4</b>
Mathematics	0.06805	0.00953	0.01290	0.02219	0.02346
Reading	0.04681	0.03588	0.02378	0.00868	0.00834
Science	0.08414	0.01687	0.01342	0.02351	0.03078
Writing	0.04262	0.02086	0.01877	0.02115	0.02257

**Table 12: False Positive Classification**

<b>Content Area</b>	<b>Overall False Positive</b>	<b>Cut 1</b>	<b>Cut 2</b>	<b>Cut 3</b>	<b>Cut 4</b>
Mathematics	0.08101	0.02173	0.02964	0.01372	0.01610
Reading	0.02840	0.01149	0.00545	0.01709	0.01769
Science	0.07029	0.01406	0.02146	0.01774	0.01739
Writing	0.01413	0.00816	0.00742	0.00543	0.00466

## 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: 2009 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 section details the critical 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 less valid 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 reference 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 HS16 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	-1.6040	2.04	0.50
2	OE	0.3461	1.31	0.68
3	OE	0.4666	1.20	0.65
4	OE	0.2525	1.34	0.66
5	GR	-1.3601	0.71	0.49
6	GR	0.2743	0.45	0.55
7	GR	0.0941	0.48	0.58
8	GR	0.6694	0.38	0.52
9	GR	0.0298	0.49	0.62
10	GR	0.8960	0.34	0.47
11	GR	1.2992	0.28	0.40
12	GR	0.6805	0.38	0.57
13	GR	0.9419	0.33	0.45
14	GR	-0.7499	0.62	0.59
15	GR	0.5060	0.41	0.62
16	GR	0.2533	0.45	0.62
17	OE	0.0727	1.38	0.61
18	OE	0.6576	1.12	0.59
19	OE	0.3300	1.29	0.68
20	OE	0.1055	1.41	0.72
21	GR	-2.8358	0.87	0.32
22	GR	0.2323	0.45	0.58
23	GR	-0.2089	0.53	0.44
24	GR	0.1163	0.47	0.67
25	GR	-0.4666	0.57	0.64
26	GR	-0.0275	0.50	0.55
27	GR	-0.3204	0.55	0.61
28	GR	0.0932	0.48	0.60
29	GR	-0.4829	0.58	0.64

<b>Item</b>	<b>Type</b>	<b>Rasch</b>	<b>PC/Mean</b>	<b>RPB/Corr</b>
30	GR	1.3458	0.27	0.53
31	GR	0.1281	0.47	0.68
32	GR	-1.3576	0.71	0.45

## Science HS16 Item Analysis

### Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

### 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	-0.0847	1.89	0.51
2	OE	0.0818	1.72	0.60
3	OE	0.3262	1.58	0.59
4	MC	-1.4873	0.83	0.39
5	MC	-0.9562	0.76	0.36
6	MC	-0.6586	0.71	0.42
7	MC	-0.8751	0.74	0.48
8	MC	0.3317	0.52	0.34
9	MC	0.4411	0.50	0.42
10	MC	-0.5669	0.69	0.48
11	MC	0.4363	0.50	0.29
12	MC	0.3819	0.51	0.41
13	MC	0.8981	0.40	0.40
14	MC	1.0067	0.38	0.43
15	MC	-0.0621	0.60	0.41
16	MC	-0.4101	0.66	0.35
17	MC	-0.5917	0.70	0.35
18	MC	0.5102	0.48	0.38
19	MC	-0.1917	0.62	0.44
20	MC	0.3460	0.52	0.45
21	MC	-0.0810	0.60	0.36
22	MC	1.0135	0.38	0.36
23	MC	0.3942	0.51	0.49
24	MC	0.1630	0.55	0.33
25	MC	-0.5584	0.69	0.47
26	MC	-0.6101	0.70	0.46
27	MC	-0.0350	0.59	0.57
28	MC	0.2548	0.53	0.37
29	MC	0.4404	0.50	0.39
30	MC	-0.4246	0.67	0.50
31	MC	-0.6901	0.71	0.40

Item	Type	Rasch	PC/Mean	RPB/Corr
32	OE	-0.0110	1.79	0.61
33	OE	0.1905	1.66	0.54
34	MC	-0.8413	0.74	0.28
35	MC	-0.0075	0.59	0.33
36	MC	-0.3601	0.65	0.52
37	MC	-1.3230	0.81	0.40
38	MC	0.5938	0.46	0.40
39	MC	-0.3564	0.65	0.42
40	MC	-0.4030	0.66	0.44
41	MC	-0.3080	0.64	0.47
42	MC	0.1359	0.56	0.43
43	MC	0.8291	0.42	0.39
44	MC	0.0766	0.57	0.48
45	MC	-0.4353	0.67	0.36
46	MC	-0.0368	0.59	0.28
47	MC	-0.2142	0.63	0.31
48	MC	-1.0201	0.77	0.46
49	MC	0.6172	0.46	0.21
50	MC	-1.4709	0.83	0.45
51	MC	0.2497	0.53	0.33
52	MC	-0.2748	0.64	0.37
53	MC	0.0360	0.58	0.38
54	MC	1.0670	0.37	0.30
55	MC	-0.9191	0.75	0.53
56	MC	-0.0013	0.58	0.45
57	MC	0.2323	0.54	0.49
58	MC	0.6652	0.45	0.32
59	MC	0.0885	0.57	0.30
60	MC	0.0739	0.57	0.53
61	MC	1.5098	0.29	0.32
62	MC	-0.2711	0.64	0.47
63	MC	0.3134	0.52	0.47
64	MC	0.0905	0.57	0.50
65	MC	0.2070	0.54	0.35

## Reading for Information HS16 Item Analysis

### Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

### 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	-0.9519	0.73	0.41
2	MC	-2.4330	0.90	0.31
3	MC	-1.0118	0.74	0.41
4	MC	-1.5425	0.81	0.43
5	OE	0.5097	0.96	0.57
6	OE	0.6873	0.92	0.48
7	MC	0.9721	0.39	0.28
8	MC	-0.0386	0.58	0.37
9	MC	-0.3423	0.63	0.40
10	MC	-1.1422	0.76	0.24
11	OE	-0.3833	1.18	0.53
12	OE	1.3093	0.76	0.56
13	MC	-1.4689	0.81	0.42
14	MC	0.8493	0.41	0.30
15	MC	-0.3638	0.64	0.24
16	MC	-0.6850	0.69	0.42
17	OE	1.0178	0.83	0.49
18	OE	2.0276	0.46	0.41

### Editing and Revising HS16 Item Analysis

**Multiple-choice Items**

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

Item	Type	Rasch	PC	RPB
1	MC	0.7217	0.61	0.21
2	MC	-0.7174	0.84	0.23
3	MC	-1.4069	0.91	0.35
4	MC	-1.3103	0.90	0.29
5	MC	0.5253	0.65	0.39
6	MC	0.2052	0.71	0.34
7	MC	-1.7882	0.93	0.27
8	MC	1.5191	0.45	0.23
9	MC	1.2878	0.50	0.35
10	MC	0.4733	0.66	0.38
11	MC	-1.0950	0.88	0.33
12	MC	1.5793	0.44	0.20
13	MC	-1.2803	0.90	0.27
14	MC	-0.8546	0.85	0.37
15	MC	0.4439	0.66	0.22
16	MC	1.0996	0.54	0.28
17	MC	0.6605	0.62	0.36
18	MC	-0.5298	0.82	0.39

### Response to Literature and Interdisciplinary Writing HS16 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	0.4270	7.11	0.62	0.01	0.01	0.06	0.07	0.21	0.19	0.27	0.12	0.05	0.01	0.00
IW1	EX	1.0651	7.85	0.72	0.02	0.02	0.03	0.04	0.10	0.11	0.30	0.17	0.14	0.05	0.02
IW2	EX	0.9835	7.89	0.71	0.02	0.01	0.04	0.05	0.09	0.14	0.25	0.19	0.13	0.05	0.02

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

**Raw, Theta, and Scale Scores for Mathematics HS16**

<b>Raw Score</b>	<b>Theta</b>	<b>Scale Score</b>
0	-6.0198	100
1	-4.6057	128
2	-3.6719	154
3	-3.0875	170
4	-2.6620	182
5	-2.3280	192
6	-2.0530	199
7	-1.8192	206
8	-1.6158	212
9	-1.4356	217
10	-1.2737	221
11	-1.1267	225
12	-0.9919	229
13	-0.8672	233
14	-0.7513	236
15	-0.6425	239
16	-0.5400	242

<b>Raw Score</b>	<b>Theta</b>	<b>Scale Score</b>
17	-0.4429	244
18	-0.3503	247
19	-0.2615	249
20	-0.1760	252
21	-0.0931	254
22	-0.0125	256
23	0.0666	259
24	0.1444	261
25	0.2216	263
26	0.2984	265
27	0.3753	267
28	0.4528	270
29	0.5314	272
30	0.6116	274
31	0.6938	276
32	0.7787	279
33	0.8668	281

<b>Raw Score</b>	<b>Theta</b>	<b>Scale Score</b>
34	0.9590	284
35	1.0560	286
36	1.1588	289
37	1.2686	292
38	1.3867	296
39	1.5152	299
40	1.6562	303
41	1.8134	308
42	1.9914	313
43	2.1978	318
44	2.4452	325
45	2.7574	334
46	3.1874	346
47	3.9051	366
48	5.1263	400

**Raw, Theta, and Scale Scores for Science HS16**

Raw Score	Theta	Scale Score
0	-5.6117	100
1	-4.3939	100
2	-3.6809	100
3	-3.2556	100
4	-2.9481	100
5	-2.7052	100
6	-2.5031	107
7	-2.3291	115
8	-2.1758	122
9	-2.0381	129
10	-1.9128	135
11	-1.7975	141
12	-1.6905	146
13	-1.5903	151
14	-1.4960	156
15	-1.4068	160
16	-1.3219	164
17	-1.2408	168
18	-1.1631	172
19	-1.0884	175
20	-1.0162	179
21	-0.9465	182
22	-0.8788	186
23	-0.8131	189
24	-0.7490	192
25	-0.6865	195

Raw Score	Theta	Scale Score
26	-0.6253	198
27	-0.5654	201
28	-0.5065	204
29	-0.4487	206
30	-0.3916	209
31	-0.3354	212
32	-0.2797	215
33	-0.2246	217
34	-0.1699	220
35	-0.1156	223
36	-0.0615	225
37	-0.0076	228
38	0.0463	231
39	0.1003	233
40	0.1543	236
41	0.2086	238
42	0.2632	241
43	0.3183	244
44	0.3738	246
45	0.4301	249
46	0.4870	252
47	0.5448	255
48	0.6036	258
49	0.6634	261
50	0.7244	264
51	0.7869	267

Raw Score	Theta	Scale Score
52	0.8508	270
53	0.9165	273
54	0.9841	276
55	1.0538	280
56	1.1258	283
57	1.2005	287
58	1.2781	290
59	1.3591	294
60	1.4439	298
61	1.5331	303
62	1.6273	307
63	1.7273	312
64	1.8342	317
65	1.9494	323
66	2.0745	329
67	2.2121	336
68	2.3653	343
69	2.5390	352
70	2.7409	362
71	2.9836	373
72	3.2907	388
73	3.7157	400
74	4.4282	400
75	5.6457	400

**Raw, Theta, and Scale Scores for Reading HS16**

Raw Score	Theta	Scale Score
0	-5.5447	100
1	-4.4302	100
2	-3.8075	100
3	-3.4328	101
4	-3.1498	112
5	-2.9116	121
6	-2.6985	128
7	-2.5006	136
8	-2.3133	143
9	-2.1339	149
10	-1.9613	155
11	-1.7946	162
12	-1.6330	167
13	-1.4759	173
14	-1.3223	179
15	-1.1712	184
16	-1.0219	190

Raw Score	Theta	Scale Score
17	-0.8735	195
18	-0.7254	201
19	-0.5772	206
20	-0.4286	212
21	-0.2791	217
22	-0.1286	223
23	0.0235	228
24	0.1777	234
25	0.3346	240
26	0.4948	245
27	0.6593	251
28	0.8286	258
29	1.0032	264
30	1.1835	271
31	1.3693	277
32	1.5599	284
33	1.7547	292

Raw Score	Theta	Scale Score
34	1.9531	299
35	2.1549	306
36	2.3601	314
37	2.5700	321
38	2.7859	329
39	3.0098	338
40	3.2441	346
41	3.4911	355
42	3.7537	365
43	4.0368	375
44	4.3495	387
45	4.7121	400
46	5.1737	400
47	5.8986	400
48	7.1042	400

**Raw, Theta, and Scale Scores for Writing HS16**

Raw Score	Theta	Scale Score
0	-4.8913	100
1	-3.6563	100
2	-2.9273	100
3	-2.4955	109
4	-2.1894	121
5	-1.9542	131
6	-1.7643	138
7	-1.6054	145
8	-1.4688	150
9	-1.3487	155
10	-1.2412	159
11	-1.1432	163
12	-1.0527	167
13	-0.9680	170
14	-0.8880	173
15	-0.8115	177
16	-0.7378	180
17	-0.6661	182
18	-0.5959	185
19	-0.5268	188
20	-0.4581	191

Raw Score	Theta	Scale Score
21	-0.3896	194
22	-0.3208	196
23	-0.2514	199
24	-0.1811	202
25	-0.1096	205
26	-0.0364	208
27	0.0387	211
28	0.1159	214
29	0.1958	217
30	0.2786	220
31	0.3646	224
32	0.4543	227
33	0.5481	231
34	0.6462	235
35	0.7490	239
36	0.8567	244
37	0.9695	248
38	1.0873	253
39	1.2101	258
40	1.3374	263
41	1.4689	268

Raw Score	Theta	Scale Score
42	1.6042	274
43	1.7432	279
44	1.8855	285
45	2.0312	291
46	2.1808	297
47	2.3348	303
48	2.4941	310
49	2.6598	316
50	2.8332	323
51	3.0156	331
52	3.2091	338
53	3.4160	347
54	3.6404	356
55	3.8884	366
56	4.1715	377
57	4.5120	391
58	4.9614	400
59	5.6876	400
60	6.9062	400