

The Connecticut Academic Performance Test: Technical Report

**Prepared by
Irene Hendrawan & Arianto Wibowo**

February 2012



Table of Contents

Table of Contents	i
List of Charts	ii
List of Tables	iii
Part 1: Introduction	1
1.1. General description of CAPT.....	1
1.2. 2011 CAPT Test Design.....	1
1.3. 2011 CAPT Test Forms.....	2
Part 2: Test Development	3
Part 3: Item Level Statistics	4
Part 4: Scaling and Equating	5
4.1. 2011 CAPT Linking Items.....	5
4.2. Calibration Process.....	6
Part 5: Test Statistics	9
5.1. Reliability.....	9
5.2. Classification Consistency and Accuracy.....	9
Part 6: CAPT3 Standards	11
Part 7: Validity	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
References	16
Appendix A: Item Analysis	17
Appendix B: Raw, Theta, and Scale Scores	23

List of Charts

Chart 1: Calibration Design for 2011 Mathematics	7
Chart 2: Calibration Design for 2011 Science	7
Chart 3: Calibration Design for 2011 Reading	7
Chart 4: Calibration Design for 2011 Writing	7

List of Tables

Table 1: 2011 CAPT Operational Test Design	1
Table 2: Summary of Item Analysis Form HS18	4
Table 3: 2011 Embedded Linking Items	5
Table 4: 2011 CAPT Equating Constants	7
Table 5: Summary of Weighting for Reading and Writing	8
Table 6: Scaling Coefficients from Base Year (CAPT2)	8
Table 7: CAPT Cronbach’s Alpha	9
Table 8: CAPT Scale Score Summary Statistics	9
Table 9: Classification Consistency	9
Table 10: Classification Accuracy	10
Table 11: False Negative Classification	10
Table 12: False Positive Classification	10
Table 13: 2011 CAPT Achievement Levels and Scale Score Ranges	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 2011 administration was the fifth operational (OP) administration of CAPT3.

1.2. 2011 CAPT Test Design

The spring 2011 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: 2011 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. 2011 CAPT Test Forms

In the 2011 administration, two main forms were available for administration: Form HS18, 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 (HS18 or HS0) was given the corresponding test session or subtest (HS18 or HS0).

Part 2: Test Development

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

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

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

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

Part 3: Item Level Statistics

Table 2 and Appendix A present a summary and detailed result of item analysis (item quality) data for Mathematics, Science, Reading and Writing, respectively. 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 HS18

Subject	Rasch		P-value		Point Biserial	
	Mean	Std	Mean	Std	Mean	Std
Mathematics	0.04	0.76	0.74	0.48	0.56	0.09
Reading for Information	-0.01	0.96	0.70	0.11	0.41	0.12
Response to Literature	0.52		7.13		0.63	
Editing and Revising	0.02	0.54	0.75	0.08	0.32	0.07
Interdisciplinary Writing	1.45	0.10	7.51	0.07	0.71	0.00
Science	0.05	0.60	0.70	0.33	0.40	0.10

Part 4: Scaling and Equating

4.1 2011 CAPT Linking Items

The 2011 CAPT Mathematics, Science, Reading for Information, and Editing & Revising tests were equated with the 2010 CAPT (HS17) subtests by embedding linking items. Linking items were counted toward students' scores.

The Live form of the 2011 CAPT (HS18) 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 2010 and 2009 tests.

Table 3: 2011 Embedded Linking Items

Content Area	Form HS18 Item Position	Form HS17 Item Position	Rasch Form HS17
Mathematics	9	9	1.0325
	10	10	0.8375
	11	11	1.0380
	12	12	-1.2969
	13	13	-0.8934
	14	14	-0.4307
	21	21	0.0743
	22	22	0.4408
	24	24	-0.2140
	27	27	-0.8362
	29	29	0.8172
	31	31	-0.7778
Science	4	4	-0.9130
	6	6	-0.9816
	7	7	-0.2726
	14	14	0.3816
	15	15	0.1785
	16	16	-0.0427
	40	40	-0.1777
	50	50	0.0604
	54	54	-0.4500

Content Area	Form HS18 Item Position	Form HS17 Item Position	Rasch Form HS17
	58	58	0.4387
	59	59	0.4366
	60	60	0.9920
	61	61	-0.2420
	62	62	-0.8544
	65	65	-0.4580
Reading	1	1	-1.2203
	2	2	-0.8614
	3	3	-0.3389
	4	4	-1.1462
	5	5	-0.2121
	6	6	1.2687
Writing	13	13	-0.2671
	14	14	-0.6365
	15	15	0.7602
	16	16	-0.5061
	17	17	0.5487
	18	18	0.7345

4.2. Calibration Process

The CAPT 2011 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 HS17 (CAPT 2010 Live). The equating was accomplished in the following steps:

1. For every content area, calibrate all items in 2011 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 2011 Mathematics

Form HS18	HS18_MA1	HS18_MA2
-----------	----------	----------

Note:

HS18_MA1 = Form HS18 Math Session 1

HS18_MA2 = Form HS18 Math Session 2

Chart 2: Calibration Design for 2011 Science

Form HS18	HS18_SC1	HS18_SC2
-----------	----------	----------

Note:

HS18_SC1 = Form HS18 Science Session 1

HS18_SC2 = Form HS18 Science Session 2

Chart 3: Calibration Design for 2011 Reading

Form HS18	HS18_RI	HS18_RL
-----------	---------	---------

Note:

HS18_RI = Form HS18 Reading for Information

HS18_RL = Form HS18 Response to Literature

Chart 4: Calibration Design for 2011 Writing

HS18	HS18_ER	HS18_IW1	HS18_IW2
------	---------	----------	----------

Note:

HS18_ER = Form HS18 Editing & Revising

HS18_IW1 = Form HS18 Interdisciplinary Writing 1

HS18_IW2 = Form HS18 Interdisciplinary Writing 2

2. Select the items linking HS18 (2011 live test) and HS17 (2010 live test). Do anchor evaluation using .3 rule between the estimates of difficulties from Step 1 and HS17 values (see Table 3 for the Rasch values of linking items between Form 18 and Form 17). 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 HS17 and Step 1 yields the equating constant. Table 4 shows the equating constants for Form 18 and Form 17.

Table 4: 2011 CAPT Equating Constants

Content Area	Equating Constant
Mathematics	0.0375
Reading	0.0132
Science	0.0533
Writing	0.1634

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 HS17 and mean of difficulty estimates of the linking items on HS18, 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 HS18. 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
HS18	0.94	0.83	0.80	0.93

Table 8: CAPT Scale Score Summary Statistics

Subject	Mean	Standard Deviation
Mathematics	255.59	45.87
Reading	244.78	45.50
Writing	264.86	49.18
Science	260.29	48.47

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	0.790	0.948	0.957	0.945	0.938
Reading	0.893	0.919	0.964	0.963	0.964
Science	0.790	0.957	0.950	0.943	0.933
Writing	0.841	0.900	0.964	0.963	0.964

Table 10: Classification Accuracy

Content Area	Overall Classification Accuracy	Cut Below Basic - Basic	Cut Basic - Proficient	Cut Proficient - Goal	Cut Goal - Advanced
Mathematics	0.850	0.963	0.970	0.961	0.956
Reading	0.896	0.939	0.961	0.975	0.973
Science	0.848	0.969	0.964	0.960	0.953
Writing	0.855	0.925	0.969	0.969	0.966

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

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

Table 11: False Negative Classification

Content Area	Overall False Negative	Cut Below Basic - Basic	Cut Basic - Proficient	Cut Proficient - Goal	Cut Goal - Advanced
Mathematics	0.079	0.015	0.016	0.020	0.027
Reading	0.059	0.034	0.036	0.015	0.008
Science	0.080	0.018	0.017	0.020	0.027
Writing	0.075	0.030	0.025	0.004	0.030

Table 12: False Positive Classification

Content Area	Overall False Positive	Cut Below Basic - Basic	Cut Basic - Proficient	Cut Proficient - Goal	Cut Goal - Advanced
Mathematics	0.072	0.022	0.014	0.018	0.017
Reading	0.044	0.026	0.003	0.010	0.018
Science	0.072	0.013	0.019	0.020	0.021
Writing	0.069	0.043	0.004	0.025	0.002

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: 2011 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 10th grade. The respondents characterized the strands as important educational outcomes to which students would be instructed prior to testing.

7.2. Scoring Quality Assurance Procedures Undertaken during Development

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

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

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

7.3. Item Quality Analysis Undertaken During Development

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

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

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

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

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

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

For Multiple-Choice (MC) Items

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

For Constructed-Response (CR) Items

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

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

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

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

1. α = case-control estimate of risk (odds ratio)
2. β = natural log of α
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* (2nd 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* (4th 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* (3rd 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 HS18 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	-0.7990	1.96	0.60
2	OE	0.1208	1.49	0.72
3	OE	1.2120	0.88	0.64
4	OE	1.3382	0.82	0.62
5	GR	-1.2992	0.72	0.34
6	GR	0.5157	0.42	0.59
7	GR	-0.0819	0.53	0.58
8	GR	0.8834	0.36	0.60
9	GR	0.9283	0.35	0.54
10	GR	0.7021	0.39	0.57
11	GR	0.9679	0.35	0.52
12	GR	-1.3256	0.72	0.54
13	GR	-1.0590	0.69	0.58
14	GR	-0.4920	0.60	0.37
15	GR	1.1557	0.32	0.59
16	GR	0.2391	0.47	0.44
17	OE	-0.5646	1.69	0.61
18	OE	-0.4177	1.79	0.65
19	OE	0.4895	1.32	0.71
20	OE	-0.4638	1.70	0.70
21	GR	0.0520	0.50	0.52
22	GR	0.6206	0.40	0.43
23	GR	-0.7974	0.65	0.62
24	GR	-0.1087	0.53	0.47
25	GR	-0.1335	0.53	0.58
26	GR	-0.5483	0.61	0.52
27	GR	-0.6612	0.62	0.40
28	GR	0.2444	0.47	0.62

Item	Type	Rasch	PC/Mean	RPB/Corr
29	GR	0.9199	0.35	0.59
30	GR	0.5580	0.41	0.53
31	GR	-0.7535	0.64	0.58
32	GR	-0.2424	0.55	0.53

Science HS18 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	0.4586	1.62	0.42
2	OE	0.5664	1.55	0.49
3	OE	0.2971	1.73	0.61
4	MC	-1.0854	0.81	0.48
5	MC	-0.7487	0.76	0.42
6	MC	-0.8371	0.77	0.24
7	MC	-0.3658	0.69	0.30
8	MC	0.6008	0.50	0.28
9	MC	-0.9867	0.79	0.39
10	MC	-0.6259	0.74	0.37
11	MC	-0.5242	0.72	0.31
12	MC	0.6032	0.50	0.36
13	MC	0.3464	0.56	0.27
14	MC	0.4102	0.54	0.35
15	MC	0.2712	0.57	0.14
16	MC	0.0275	0.62	0.52
17	MC	-0.4565	0.71	0.34
18	MC	0.5870	0.51	0.49
19	MC	-0.6406	0.74	0.31
20	MC	-0.7968	0.77	0.49
21	MC	-0.6502	0.74	0.45
22	MC	0.6395	0.50	0.30
23	MC	0.8375	0.46	0.38
24	MC	-0.0373	0.63	0.35
25	MC	-0.1386	0.65	0.53
26	MC	0.5931	0.51	0.39
27	MC	-0.2000	0.66	0.49
28	MC	0.8366	0.46	0.48
29	MC	0.8552	0.45	0.45
30	MC	0.4695	0.53	0.38

Item	Type	Rasch	PC/Mean	RPB/Corr
31	MC	-0.3209	0.69	0.53
32	OE	0.3079	1.70	0.57
33	OE	-0.3468	2.16	0.54
34	MC	0.6097	0.50	0.41
35	MC	0.4364	0.54	0.25
36	MC	-0.0404	0.63	0.26
37	MC	-0.7038	0.75	0.33
38	MC	-0.4456	0.71	0.50
39	MC	-0.0053	0.63	0.37
40	MC	-0.4623	0.71	0.33
41	MC	0.4313	0.54	0.27
42	MC	1.1413	0.39	0.32
43	MC	0.1058	0.60	0.45
44	MC	-0.3281	0.69	0.29
45	MC	1.2958	0.36	0.27
46	MC	-0.7592	0.76	0.33
47	MC	0.9319	0.44	0.26
48	MC	-0.9027	0.78	0.52
49	MC	0.7268	0.48	0.29
50	MC	0.2164	0.58	0.37
51	MC	-0.1218	0.65	0.49
52	MC	0.6392	0.50	0.45
53	MC	0.2182	0.58	0.39
54	MC	-0.4375	0.71	0.46
55	MC	0.5328	0.52	0.46
56	MC	0.5994	0.50	0.41
57	MC	-0.1661	0.66	0.28
58	MC	0.3679	0.55	0.38
59	MC	0.4558	0.53	0.50
60	MC	1.1107	0.40	0.35
61	MC	-0.3303	0.69	0.45
62	MC	-0.7255	0.75	0.53
63	MC	-0.4821	0.71	0.51
64	MC	0.1295	0.60	0.36
65	MC	-0.5196	0.72	0.53

Reading for Information HS18 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	-1.2451	0.78	0.25
2	MC	-0.8559	0.72	0.23
3	MC	-0.3263	0.63	0.21
4	MC	-1.1559	0.76	0.37
5	OE	0.2662	1.02	0.53
6	OE	1.2853	0.72	0.53
7	MC	-0.0351	0.57	0.27
8	MC	-0.6991	0.69	0.38
9	MC	0.0589	0.55	0.29
10	MC	-0.6438	0.68	0.39
11	OE	1.0177	0.81	0.49
12	OE	0.9372	0.83	0.49
13	MC	-0.0383	0.57	0.37
14	MC	-0.7243	0.70	0.41
15	MC	-1.0285	0.74	0.46
16	MC	-0.4212	0.64	0.49
17	OE	1.6853	0.59	0.59
18	OE	1.6550	0.65	0.55

Editing and Revising HS18 Item Analysis

Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

Item	Type	Rasch	PC/Mean	RPB/Corr
1	MC	-0.7653	0.86	0.35
2	MC	-0.6779	0.85	0.20
3	MC	0.8164	0.62	0.25
4	MC	0.3617	0.70	0.41
5	MC	0.2527	0.72	0.25
6	MC	-0.0722	0.77	0.27
7	MC	0.4120	0.69	0.37
8	MC	-0.7034	0.85	0.40
9	MC	0.1722	0.73	0.23
10	MC	0.0329	0.75	0.34
11	MC	-0.4667	0.82	0.33
12	MC	0.3712	0.70	0.37
13	MC	-0.2873	0.80	0.40
14	MC	-0.6519	0.85	0.39
15	MC	0.8414	0.62	0.23
16	MC	-0.3955	0.82	0.34
17	MC	0.4802	0.68	0.34
18	MC	0.6467	0.65	0.35

Response to Literature and Interdisciplinary Writing HS18 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	0.5187	7.13	0.63	0.01	0.01	0.06	0.07	0.21	0.18	0.28	0.12	0.06	0.01	0.00
IW1	EX	1.5213	7.46	0.71	0.02	0.02	0.05	0.05	0.13	0.13	0.33	0.14	0.10	0.02	0.01
IW2	EX	1.3798	7.56	0.71	0.01	0.01	0.04	0.06	0.12	0.16	0.29	0.17	0.09	0.03	0.01

Appendix B: Raw, Theta, and Scale Scores

Raw, Theta, and Scale Scores for Mathematics HS18

Raw Score	Theta	Scale Score
0	-5.4779	104
1	-4.2139	139
2	-3.4447	161
3	-2.9725	174
4	-2.6255	184
5	-2.3485	192
6	-2.1163	198
7	-1.9155	204
8	-1.7374	209
9	-1.5766	213
10	-1.4292	217
11	-1.2925	221
12	-1.1646	225
13	-1.0437	228
14	-0.9287	232
15	-0.8186	235
16	-0.7125	238

Raw Score	Theta	Scale Score
17	-0.6099	240
18	-0.5101	243
19	-0.4127	246
20	-0.3172	249
21	-0.2231	251
22	-0.1302	254
23	-0.0380	256
24	0.0538	259
25	0.1456	262
26	0.2375	264
27	0.3302	267
28	0.4237	269
29	0.5185	272
30	0.6150	275
31	0.7135	278
32	0.8147	280
33	0.9190	283

Raw Score	Theta	Scale Score
34	1.0269	286
35	1.1394	289
36	1.2571	293
37	1.3812	296
38	1.5129	300
39	1.6536	304
40	1.8056	308
41	1.9716	313
42	2.1557	318
43	2.3649	324
44	2.6108	331
45	2.9164	339
46	3.3344	351
47	4.0347	371
48	5.2411	400

Raw, Theta, and Scale Scores for Science HS18

Raw Score	Theta	Scale Score
0	-5.6071	100
1	-4.3911	100
2	-3.6806	100
3	-3.2578	100
4	-2.9526	100
5	-2.7118	101
6	-2.5118	111
7	-2.3398	119
8	-2.1882	126
9	-2.0523	133
10	-1.9286	139
11	-1.8148	145
12	-1.7092	150
13	-1.6102	154
14	-1.5170	159
15	-1.4288	163
16	-1.3447	167
17	-1.2643	171
18	-1.1871	175
19	-1.1127	179
20	-1.0407	182
21	-0.9710	186
22	-0.9032	189
23	-0.8371	192
24	-0.7724	195
25	-0.7092	198

Raw Score	Theta	Scale Score
26	-0.6472	201
27	-0.5861	204
28	-0.5261	207
29	-0.4667	210
30	-0.4082	213
31	-0.3502	216
32	-0.2926	219
33	-0.2355	221
34	-0.1787	224
35	-0.1221	227
36	-0.0657	230
37	-0.0093	232
38	0.0470	235
39	0.1035	238
40	0.1600	241
41	0.2169	243
42	0.2741	246
43	0.3317	249
44	0.3897	252
45	0.4483	255
46	0.5076	257
47	0.5677	260
48	0.6286	263
49	0.6905	266
50	0.7536	269
51	0.8179	273

Raw Score	Theta	Scale Score
52	0.8836	276
53	0.9509	279
54	1.0200	282
55	1.0910	286
56	1.1643	289
57	1.2401	293
58	1.3187	297
59	1.4006	301
60	1.4861	305
61	1.5759	309
62	1.6707	314
63	1.7711	319
64	1.8783	324
65	1.9936	330
66	2.1188	336
67	2.2562	342
68	2.4092	350
69	2.5828	358
70	2.7843	368
71	3.0265	380
72	3.3330	395
73	3.7573	400
74	4.4691	400
75	5.6860	400

Raw, Theta, and Scale Scores for Reading HS18

Raw Score	Theta	Scale Score
0	-5.2050	100
1	-4.1438	100
2	-3.5594	102
3	-3.2084	115
4	-2.9432	125
5	-2.7202	133
6	-2.5214	140
7	-2.3377	147
8	-2.1646	153
9	-1.9995	159
10	-1.8410	165
11	-1.6882	171
12	-1.5401	176
13	-1.3960	182
14	-1.2553	187
15	-1.1170	192
16	-0.9808	197

Raw Score	Theta	Scale Score
17	-0.8460	202
18	-0.7121	207
19	-0.5788	212
20	-0.4456	216
21	-0.3120	221
22	-0.1777	226
23	-0.0417	231
24	0.0964	236
25	0.2375	241
26	0.3826	247
27	0.5325	252
28	0.6882	258
29	0.8505	264
30	1.0201	270
31	1.1971	277
32	1.3816	283
33	1.5729	290

Raw Score	Theta	Scale Score
34	1.7707	298
35	1.9744	305
36	2.1843	313
37	2.4014	321
38	2.6274	329
39	2.8642	338
40	3.1145	347
41	3.3805	357
42	3.6644	367
43	3.9694	378
44	4.3031	390
45	4.6840	400
46	5.1608	400
47	5.8968	400
48	7.1074	400

Raw, Theta, and Scale Scores for Writing HS18

Raw Score	Theta	Scale Score
0	-4.6408	100
1	-3.4469	100
2	-2.7691	101
3	-2.3786	117
4	-2.1045	128
5	-1.8932	136
6	-1.7204	143
7	-1.5736	149
8	-1.4451	154
9	-1.3301	159
10	-1.2252	163
11	-1.1281	167
12	-1.0371	171
13	-0.9509	174
14	-0.8684	178
15	-0.7889	181
16	-0.7116	184
17	-0.6360	187
18	-0.5616	190
19	-0.4880	193
20	-0.4149	196

Raw Score	Theta	Scale Score
21	-0.3418	199
22	-0.2685	202
23	-0.1946	205
24	-0.1199	208
25	-0.0441	211
26	0.0332	214
27	0.1122	217
28	0.1932	220
29	0.2765	224
30	0.3627	227
31	0.4520	231
32	0.5450	235
33	0.6422	238
34	0.7442	243
35	0.8518	247
36	0.9655	251
37	1.0863	256
38	1.2149	261
39	1.3520	267
40	1.4978	273
41	1.6527	279

Raw Score	Theta	Scale Score
42	1.8159	286
43	1.9868	293
44	2.1644	300
45	2.3480	307
46	2.5373	315
47	2.7328	323
48	2.9359	331
49	3.1482	339
50	3.3715	348
51	3.6068	358
52	3.8546	368
53	4.1144	378
54	4.3866	389
55	4.6752	400
56	4.9902	400
57	5.3534	400
58	5.8163	400
59	6.5461	400
60	7.7599	400