



# Classroom Diagnostic Tool

Test Design and CAT Configurations

7/31/2015

## TEST DESIGN AND CAT CONFIGURATIONS

The Classroom Diagnostic Tools (CDT) features a number of tests. Tests in Mathematics, Reading, and Writing have been available since April 2014 for students in grades 3 through 5. Tests in Mathematics have been available since October 2010 for students in grades 6 and above. Tests in Reading have been available since April 2011 for students in grades 6 and above. Tests in Writing have been available since October 2011 for students in grades 6 and above.

This document details the operational CDT test design and configuration of the CAT algorithm. Test design elements include the number of diagnostic categories, the number of operational items to administer per diagnostic category, and the number of embedded field test items. CAT algorithm elements include entry point, item selection criteria, test navigation, and termination.

## OPERATIONAL TEST DESIGN

### NUMBER OF DIAGNOSTIC CATEGORIES

The CDT tests include selected-response items. The Math and Writing assessments include multiple-choice items only and the Reading assessments include multiple-choice and evidence-based selected-response items. All items in the content areas of mathematics, reading, and writing will be aligned to the Missouri Learning Standards. Each CDT is broken into four or five reporting categories and the items in the pool are grouped by these categories. The categories for each of the CDT tests are listed below.

#### Mathematics Lower Grades and Mathematics

- Numbers and Operations
- Algebraic Concepts
- Geometry
- Measurement, Data, and Probability

#### Reading Lower Grades and Reading

- Key Ideas and Details-Literature Text
- Key Ideas and Details-Informational Text
- Craft and Structure, and Integration of Knowledge and Ideas-Literature Text
- Craft and Structure, and Integration of Knowledge and Ideas-Informational Text
- [Language](#): Vocabulary Acquisition and Use

#### Writing Lower Grades and Writing/English Composition

- [Text Type and Purposes](#): Quality of Writing: Focus and Organization
- [Text Type and Purposes](#): Quality of Writing: Content and Style
- [Production and Distribution](#): Quality of Writing: Editing

- **Language:** Conventions: Punctuation, Capitalization, and Spelling
- **Language:** Conventions: Grammar and Sentence Formation

## NUMBER OF ITEMS PER DIAGNOSTIC CATEGORY

There were various factors considered when determining the number of operational items to administer per category. The goal of the CDT is to provide diagnostic information. Therefore, the test must include a sufficient number of items to provide meaningful scores with low standard errors. However, testing time is limited and the item pools are finite. A very long test may produce lower standard errors, but if it is considered to be “too long” will teachers use it? Also, the longer the test, the more the items are exposed

Prior to the launch of the first operational CDTs in fall of 2010, simulations were run of various test lengths. Considering test time factors and simulation results for various test lengths, it was determined that CDT tests with four diagnostic categories would have 12–15 items per category (48–60 items total) and CDT tests with five diagnostic categories would have 10–12 items per category (50–60 items total).

## CAT ALGORITHM

This section covers elements of the CAT algorithm including entry point, item selection criteria, test navigation, and termination.

### ENTRY POINT

All CDTs other than Reading Lower Grades and Reading begin with a small “locator” section in which one or two items per diagnostic category are administered. The order of the diagnostic categories is random. The two CDTs in the reading content area are slightly different because they are passage-based. Those, too, have a small “locator” section but they may not contain one or two items for each diagnostic category because not all passages have an item for each diagnostic category.

The CAT algorithm is designed to administer items targeted for the individual student based on performance. However, student performance in the current test setting is not known at the beginning of the test. With no prior information about a student, the starting point in each diagnostic category is an item of average difficulty. For CDT tests that are not course-specific (Mathematics Lower Grades, Mathematics, Reading Lower Grades, Reading, Writing Lower Grades, and Writing/English Composition), the student’s grade is considered in selecting an item of average difficulty. For example, a grade 7 student taking CDT Mathematics will start with an item near the average difficulty of grade 7 items in the pool.

If a student has previously taken the CDT, the prior CDT scores are used to give the CAT algorithm a “head start.” In this case, the first item in each diagnostic category is selected to match the characteristics of the prior information rather than an average item. For example, if a student previously took the CDT Mathematics test and scored very high in “Measurement, Data, and Probability,” then the first item selected in that diagnostic category will be more difficult than the grade level average.

The CAT algorithm includes a randomization component when selecting items to control item exposure. That is, one item is selected from among a set of items that are near the targeted item difficulty. This is especially important at the beginning of the CDT when no prior information is available. Randomization of items and diagnostic categories ensure that students will not see the same set of items in the same order even when all of the students are assigned items of average difficulty.

## **ITEM SELECTION CRITERIA**

Once the initial set of items has been administered, the CAT algorithm is designed to administer items targeted for the individual student based on performance. In targeting items, the CAT algorithm uses Rasch ability estimates from the current test session and considers a number of factors including test blueprint, response probability, item pool refinement, and passage-related concerns. Each of these is discussed in detail on the following pages.

## **RASCH ABILITY ESTIMATES**

The CAT algorithm has access to all item parameters in the item pool. After each item response, Rasch ability estimates and standard errors are calculated via maximum likelihood estimation (MLE) for the total test and each diagnostic category. In the case of zero (all items incorrect) and perfect (all items correct) scores, a correction factor is applied before computing the relevant maximum likelihood estimates. A fractional value is added to a zero score and subtracted from a perfect score before estimation.

After the locator section of the CDT, but before a student has taken many items in each diagnostic category, the total Rasch ability estimate is used in item selection. This is because total and diagnostic category ability estimates tend to be highly correlated and the total estimate does not change as dramatically as diagnostic category estimates given one additional item. Using the total estimate at this point prevents students from experiencing extreme fluctuations in the difficulty of items.

While use of the total Rasch ability estimate makes sense early in the test, the goal of the CDT is to be diagnostic, and some students' exhibit clear strengths and areas of need in different diagnostic categories. Therefore, after four or five items have been administered in a diagnostic category, the corresponding Rasch ability estimate for that diagnostic category is used in item selection.

## **TEST BLUEPRINT**

The CAT algorithm closely resembles a modified constrained CAT (MCCAT) design (Leung, Chang, & Hau, 2003). The general idea is that the CAT algorithm is configured with an upper and lower bound that specifies the minimum and maximum number of items that will be administered to students for both total and diagnostic categories.

## **RESPONSE PROBABILITY**

No matter which Rasch ability estimate is used in selecting an item, total or diagnostic category estimate, the CAT algorithm targets items where the student has response probability (RP) of answering correctly, based on the Rasch ability estimate and item's difficulty. The most efficient way to run a CAT is to select items where RP is 0.5. That is, select items where the student has a 50%

chance of getting the item correct. This response probability produces the smallest standard error for any given number of items.

Prior to the launch of the first operational CDTs in fall of 2010, simulations were run for various response probabilities.

Based on results of the simulations, the CDTs designed for students in grade 6 and above, the response probability is set at 0.5. This is based on the desire for low standard errors at the diagnostic category level and the grade level of students testing. As part of the CDT training, students are told that the test is computer adaptive and designed to challenge them.

For CDTs designed for students in grades 3 through 5, the response probability is set at 0.65. This response probability results in higher standard errors for the same number of items. However, there was concern that younger students may not have much experience with tests designed to be so challenging and could conceivably give up on a test that is perceived to be “too hard.”

### ITEM POOL REFINEMENT

The CAT algorithm has configurable elements that allow for refinement of the item pool used in item selection. The two configurable elements are:

**Restrict pool** — The ability to restrict the available item pool by grade/course at various points in the test.

For example, no Algebra I items will be administered in the first 5 items.

**Favor items** — The ability to favor items that are close to the student’s grade when evaluating items near a student’s estimated score.

For example, if a student is in grade 8 and the item selection routine finds appropriate items (in terms of difficulty) in grades 4, 5, 6, 7, and 8, item selection can favor items at or close to grade 8. It is possible that no items near a student’s grade are appropriate in terms of difficulty. In such a case, the CAT algorithm will select items further away from the student’s grade, but appropriate based on item difficulty.

The difference between restricting the pool and favoring items is that when the pool is restricted, some items may NOT be selected. With favoring, all non-restricted items are eligible for administration, but they are made more or less LIKELY to be selected based on closeness to student grade

### PASSAGE RELATED CONCERNS

As previously mentioned, the CDTs in the reading content area are passage-based. CDT passages have between one and seven associated items. The CAT algorithm does not require that all items associated with a passage be administered. Instead, it evaluates all possible combinations of items within a passage. Item sequencing within a passage is preserved when items are presented to the student. For example, if a six-item passage is selected and items 1 and 4 are NOT administered, then the items administered in order will be 2, 3, 5, and 6.

The configurable elements of passage-based CAT include:

**Passage minimum percent** — Define the minimum percentage of the items associated with a passage to be used.

For example, if the passage minimum percent is set at 80, then the selection routine will consider combinations such as 1 of 1 (100%), 4 of 5 (80%), 5 of 6 (83%), and 6 of 6 (100%). It will not consider combinations such as 1 of 2 (50%), 3 of 4 (75%), 3 of 5 (60%), etc. Near the end of a test, the passage minimum percent constraint may need to be loosened in order to meet content constraints such as number of items per diagnostic category.

**Passage evaluation criteria** — Multiple factors are considered when evaluating and ranking each passage combination to determine the best combination to administer to a student. They include:

Percent of items associated with the passage used; the higher the percent, the higher the combination is ranked

Number of items associated with the passage used; the higher the number, the higher the combination is ranked

Distance between items' difficulties and student's estimated score; the smaller the distance, the higher the combination is ranked

Distance between the items' grade levels and the student's grade level; the smaller the distance, the higher the combination is ranked

Different weights may be assigned to each of the factors. For example, if all of the weight is put on number of items used, then the algorithm will select the passages with the most associated items and administer all of them until the maximum number of items is reached.

## TEST NAVIGATION

Currently all CDT tests except Reading Lower Grades and Reading do not allow skipping items or backing up and changing answers. On CDTs in the reading content area, students are allowed to skip items within a passage. For example, when presented with a passage and five associated items, the student does not have to answer questions one through five in that order without skipping. If a student tries to navigate to the next passage without answering all of the items associated with a passage, the test engine will prompt the student to answer all items and will not move on to the next passage until all are answered.

## TERMINATION

The CAT algorithm allows for both a fixed- or variable-length test.

With fixed length, the test ends when a student has taken a pre-defined number of items total and in each diagnostic category.

With variable length, the algorithm stops administering items from a diagnostic category when one of two conditions is satisfied:

A student has taken at least a pre-defined minimum number of items in that diagnostic category

and the standard error is below a pre-defined threshold

OR

A student has taken a pre-defined maximum number of items in that diagnostic category

The test ends when one of the two conditions above is satisfied for each of the diagnostic categories.

Note with both fixed- and variable-length tests, there is no requirement that the pre-defined number of items in diagnostic categories be equal.

## TEST LENGTH

Each CDT should take the typical student 45 to 90 minutes to complete. Each CDT is between 45 and 60 items in length. Districts and schools may elect to administer an entire CDT at once or spread testing over two consecutive days. It is highly recommended that all grades 3-5 testing be administered over multiple days. It is also recommended that the Reading CDT be administered over two consecutive days due to the additional reading load.

## CAT CONFIGURATION – MATHEMATICS LOWER GRADES

The test has four diagnostic categories. Each student will take between 12 and 15 operational items per diagnostic category for a total test of 48 to 60 operational items. With no prior information about a student, the starting point in each diagnostic category will be an item of average difficulty by grade level. For example, a grade 4 student will start with an item near the average difficulty of grade 4 items. Items are selected where the response probability is 0.65, meaning a student has a 65% chance of answering correctly. The CAT algorithm will stop administering items in a diagnostic category when one of two conditions is satisfied:

A student has taken at least 12 operational items in that diagnostic category and the standard error is below 0.62

A student has taken 15 operational items in that diagnostic category

Functionality is used to restrict the pool and to favor items close to a student's grade. The pool restrictions are:

No grade 7 items will be administered in the first 5 items

No grade 8 items will be administered in the first 10 items

No Algebra I items will be administered in the first 20 items

No Geometry or Algebra II items will be administered

Simulations were run with this configuration. On average:

A total of 52 operational items are administered – about 13 per diagnostic category

Standard error for the total score is 0.30

Standard errors for the diagnostic categories are in the range of 0.61 to 0.63

### CAT Configuration Summary – Mathematics

	Lower Grades	Mathematics
<b>Number of DCs</b>	4	4
<b>Number of OP Items per DC</b>	12–15	12–15
<b>Number of OP Items Total</b>	48–60	48–60
<b>Number of FT Items Total</b>	5	5
<b>Entry Point</b>		
No Prior CDT	average item by grade	average item by grade
Prior CDT	prior diagnostic scores	prior diagnostic scores
<b>Item selection</b>		
Rasch Ability Estimates	After locator, use total estimate until the fifth item in a DC; then switch to DC estimate	After locator, use total estimate until the fifth item in a DC; then switch to DC estimate
Response Probability	0.65	0.50
Favor Items	close to student grade	close to student grade
Pool Restrictions	Items 1–5: no Grade 7 Items 1–10: no Grade 8 Items 1–20: no Algebra I No Geometry No Algebra II	Items 1–5: no Algebra I Items 1–10: no Geometry Items 1–20: no Algebra II
<b>Navigation</b>	no skip; no backtrack	no skip; no backtrack
<b>Termination</b>	12 items per DC, SE < 0.62 OR 15 items per DC	12 items per DC, SE < 0.60 OR 15 items per DC

DC = Diagnostic Category

### CAT CONFIGURATION – MATHEMATICS

The test has four diagnostic categories. Each student will take between 12 and 15 operational items per diagnostic category for a total test of 48 to 60 operational items. Tests also included five field test items for students in grade 6 only. With no prior information about a student, the starting point in each diagnostic category will be an item of average difficulty by grade level. For example, a grade 7 student will start with an item near the average difficulty of grade 7 items. Items are selected where the response probability is 0.5, meaning a student has a 50% chance of answering correctly. The CAT algorithm will stop administering items in a diagnostic category when one of two conditions is satisfied:

A student has taken at least 12 operational items in that diagnostic category and the standard error is below 0.60

A student has taken 15 operational items in that diagnostic category

Functionality is used to restrict the pool and to favor items close to a student's grade. The pool restrictions are:

No Algebra I items will be administered in the first 5 items

No Geometry items will be administered in the first 10 items

No Algebra II items will be administered in the first 20 items

Simulations were run with this configuration. On average:

A total of 52 operational items are administered – about 13 per diagnostic category

Standard error for the total score is 0.30

Standard errors for the diagnostic categories are in the range of 0.60 to 0.63

## CAT CONFIGURATION – READING LOWER GRADES

The test has five diagnostic categories. Each student will take between 10 and 12 operational items per diagnostic category for a total test of 50 to 60 operational items. With no prior information about a student, the starting point in each diagnostic category will be an item of average difficulty by grade level. For example, a grade 4 student will start with an item near the average difficulty of grade 4 items. Items are selected where the response probability is 0.65, meaning a student has a 65% chance of answering correctly. The CAT algorithm will stop administering items in a diagnostic category when one of two conditions is satisfied:

A student has taken at least 10 operational items in that diagnostic category and the standard error is below 0.77

A student has taken 12 operational items in that diagnostic category

Functionality is used to run CAT with passages and favor items close to student's grade. There are no pool restrictions.

Passage minimum percent is set at 66%. That is, whenever possible, only passage combinations that use 66% or more of the associated items are used. (Near the end of a test, the passage minimum percent constraint may need to be loosened in order to meet content constraints.) Many simulations were run to arrive at this percent. On the one hand, we want to minimize testing time and reading load. Therefore, we do not want students reading long passages for only one or two items. On the other hand, using all items associated with a passage may not be desirable since some items are far from a student's estimated score. Given a limited number of items, we don't want to take up spots with items that are either too easy or too hard.

In evaluating and ranking passages, percent of items associated with the passage is not used. Simulation results indicate that if it is factored into evaluations, students take many short passages because 1 of 1 (100%) and 2 of 2 (100%) are ranked higher than 5 of 6 (83%) and 4 of 5 (80%), for example.

Simulations were run with this configuration. On average:

A total of 55 operational items are administered – about 11 per diagnostic category

A total of 16 passages are administered

Standard error for the total score is 0.31

Standard errors for the diagnostic categories are in the range of 0.75 to 0.80

Note that the standard error is higher for in reading than the other content areas. This is because Reading/ and Reading Lower Grades are passage-based. Rather than selecting one targeted item at a time, the item selection routine evaluates and selects multiple items associated with a given passage. In general, items selected in this manner are not as close to the targeted response probability as stand-alone items selected one by one.

## CAT CONFIGURATION – READING

The test has five diagnostic categories. Each student will take between 10 and 12 operational items per diagnostic category for a total test of 50 to 60 operational items. Tests also included five to seven field test items for students in grade 6 only. With no prior information about a student, the starting point in each diagnostic category will be an item of average difficulty by grade level. For example, a grade 7 student will start with an item near the average difficulty of grade 7 items. Items are selected where the response probability is 0.5, meaning a student has a 50% chance of answering correctly. The CAT algorithm will stop administering items in a diagnostic category when one of two conditions is satisfied:

A student has taken at least 10 operational items in that diagnostic category and the standard error is below 0.75

A student has taken 12 operational items in that diagnostic category

Functionality is used to run CAT with passages and favor items close to student's grade. There are no pool restrictions.

Passage minimum percent is set at 66%. That is, whenever possible, only passage combinations that use 66% or more of the associated items are used. (Near the end of a test, the passage minimum percent constraint may need to be loosened in order to meet content constraints.) Many simulations were run to arrive at this percent. On the one hand, testing time and reading load should be minimized. Therefore, students should not have to read long passages for only one or two items. On the other hand, using all items associated with a passage may not be desirable since some items are far from a student's estimated score. Given a limited number of items, those that are either too easy or too hard should not be used.

In evaluating and ranking passages, percent of items associated with the passage is not used. Simulation results indicate that if it is factored into evaluations, students take many short passages because 1 of 1 (100%) and 2 of 2 (100%) are ranked higher than 5 of 6 (83%) and 4 of 5 (80%), for example.

Simulations were run with this configuration. On average:

A total of 55 operational items are administered – about 11 per diagnostic category

A total of 16 passages are administered

Standard error for the total score is 0.31

Standard errors for the diagnostic categories are in the range of 0.74 to 0.78

### CAT Configuration Summary – Reading

	Reading Lower Grades	Reading
<b>Number of DCs</b>	5	5
<b>Number of OP Items per DC</b>	10–12	10–12
<b>Number of OP Items Total</b>	50–60	50–60
<b>Number of FT Items Total</b>	2 (EBSRs only)	1 FT passage (either 5 MC and 1 EBSR or 6 MC)
<b>Entry Point</b>		
No Prior CDT	average item by grade	average item by grade
Prior CDT	prior diagnostic scores	prior diagnostic scores
<b>Item selection</b>		
Rasch Ability Estimates	After locator, use total estimate until the fifth item in a DC; then switch to DC estimate	After locator, use total estimate until the fifth item in a DC; then switch to DC estimate
Response Probability	0.65	0.50
Favor Items	close to student grade	close to student grade
Pool Restrictions	None	None
Passage Min %	66	66
<b>Navigation</b>	skip items within passage	skip items within passage
<b>Termination</b>	10 items per DC, SE < 0.77 OR 12 items per DC	10 items per DC, SE < 0.75 OR 12 items per DC

DC = Diagnostic Category

### CAT CONFIGURATION – WRITING LOWER GRADES

The test has five diagnostic categories. Each student will take between 10 and 12 operational items per diagnostic category for a total test of 50 to 60 operational items. With no prior information about a student, the starting point in each diagnostic category will be an item of average difficulty by grade level. For example, a grade 4 student will start with an item near the average difficulty of grade 4 items. Items are selected where the response probability is 0.65, meaning a student has a 65% chance of

answering correctly. The CAT algorithm will stop administering items in a diagnostic category when one of two conditions is satisfied:

A student has taken at least 10 operational items in that diagnostic category and the standard error is below 0.67

A student has taken 12 operational items in that diagnostic category

Functionality is used to favor items close to the student's grade. There are no pool restrictions.

Simulations were run with this configuration. On average:

A total of 55 operational items are administered – about 11 per diagnostic category

Standard error for the total score is 0.30

Standard errors for the diagnostic categories are in the range of 0.67 to 0.69

### **CAT CONFIGURATION – WRITING**

The test has five diagnostic categories. Each student will take between 10 and 12 operational items per diagnostic category for a total test of 50 to 60 operational items. Tests also included five field test items for students in grade 6 only. With no prior information about a student, the starting point in each diagnostic category will be an item of average difficulty by grade level. For example, a grade 7 student will start with an item near the average difficulty of grade 7 items. Items are selected where the response probability is 0.5, meaning a student has a 50% chance of answering correctly. The CAT algorithm will stop administering items in a diagnostic category when one of two conditions is satisfied:

A student has taken at least 10 operational items in that diagnostic category and the standard error is below 0.65

A student has taken 12 operational items in that diagnostic category

Functionality is used to favor items close to the student's grade. There are no pool restrictions.

Simulations were run with this configuration. On average:

A total of 55 to 56 operational items are administered – about 11 per diagnostic category

Standard error for the total score is 0.29

Standard errors for the diagnostic categories are in the range of 0.65 to 0.68

### CAT Configuration Summary – Writing

	Writing Lower Grades	Writing/
<b>Number of DCs</b>	5	5
<b>Number of OP Items per DC</b>	10–12	10–12
<b>Number of OP Items Total</b>	50–60	50–60
<b>Number of FT Items Total</b>	0	5 (grade 6 students only)
<b>Entry Point</b>		
No Prior CDT	average item by grade	average item by grade
Prior CDT	prior diagnostic scores	prior diagnostic scores
<b>Item selection</b>		
Rasch Ability Estimates	After locator, use total estimate until the fifth item in a DC; then switch to DC estimate	After locator, use total estimate until the fifth item in a DC; then switch to DC estimate
Response Probability	0.65	0.50
Favor Items	close to student grade	close to student grade
Pool Restrictions	None	None
<b>Navigation</b>	no skip; no backtrack	no skip; no backtrack
<b>Termination</b>	10 items per DC, SE < 0.67 OR 12 items per DC	10 items per DC, SE < 0.65 OR 12 items per DC

DC = Diagnostic Category