

MontCAS
Criterion-Referenced Test
(Montana CRT)
2013–14
Technical Report

TABLE OF CONTENTS

CHAPTER 1	OVERVIEW OF THE MONTANA CRITERION-REFERENCED TEST	1
1.1	PURPOSE OF THE ASSESSMENT SYSTEM	1
1.2	PURPOSE OF THIS REPORT	2
CHAPTER 2	ASSESSMENT AND TEST DEVELOPMENT PROCESS.....	3
2.1	TEST SPECIFICATIONS	3
2.1.1	Criterion-Referenced Test.....	3
2.1.2	Item Types	3
2.1.3	Description of Test Design	3
2.2	SCIENCE TEST SPECIFICATIONS	4
2.2.1	Standards.....	4
2.2.2	Item Types	4
2.2.3	Test Design	4
2.2.4	Blueprints (Distribution of Points across Standards)	5
2.2.5	Depth of Knowledge.....	5
2.2.6	Use of Calculators and Reference Sheets.....	6
2.3	TEST DEVELOPMENT PROCESS.....	6
2.3.1	Item Development	6
2.3.2	Item Reviews at Measured Progress	7
2.3.3	Item Reviews at State Level	7
2.3.4	Bias and Sensitivity Review	8
2.3.5	Reviewing and Refining	8
2.3.6	Item Editing	8
2.3.7	Item Selection and Operational Test Assembly	8
2.3.8	Operational Test Draft Review.....	9
2.3.9	Alternative Presentations	9
2.4	TEST SESSIONS	9
CHAPTER 3	TEST ADMINISTRATION	10
3.1	RESPONSIBILITY FOR ADMINISTRATION	10
3.2	ADMINISTRATION PROCEDURES	10
3.3	PARTICIPATION REQUIREMENTS AND DOCUMENTATION	11
3.3.1	Students with Disabilities	12
3.4	ADMINISTRATOR TRAINING.....	12
3.5	DOCUMENTATION OF ACCOMMODATIONS.....	12
3.6	TEST SECURITY AND ADMINISTRATION IRREGULARITIES	13
3.7	TEST ADMINISTRATION WINDOW	13
3.8	SERVICE CENTER.....	14
CHAPTER 4	SCORING.....	15
4.1	MACHINE-SCORED ITEMS	15

4.2	PERSON-SCORED ITEMS.....	16
4.2.1	Scoring Location and Staff.....	17
4.2.2	Reader Recruitment and Qualifications.....	18
4.2.3	Methodology for Scoring Polytomous Items.....	19
4.2.4	Reader Training.....	19
4.2.5	Leadership Training.....	21
4.2.6	Monitoring of Scoring Quality Control.....	22
CHAPTER 5	CLASSICAL ITEM ANALYSIS.....	26
5.1	CLASSICAL DIFFICULTY AND DISCRIMINATION INDICES.....	26
5.2	DIFFERENTIAL ITEM FUNCTIONING.....	28
5.3	DIMENSIONALITY ANALYSIS.....	29
CHAPTER 6	ITEM RESPONSE THEORY SCALING AND EQUATING.....	32
6.1	ITEM RESPONSE THEORY.....	32
6.2	ITEM RESPONSE THEORY RESULTS.....	34
6.3	EQUATING.....	35
6.4	EQUATING RESULTS.....	35
6.5	ACHIEVEMENT STANDARDS.....	36
6.5.1.	Distributions.....	36
6.6	SCALED SCORES.....	36
6.6.1	Description of Scale.....	36
6.6.2	Calculations.....	37
6.6.3	Distributions.....	38
CHAPTER 7	RELIABILITY.....	39
7.1	RELIABILITY AND STANDARD ERRORS OF MEASUREMENT.....	40
7.2	SUBGROUP RELIABILITY.....	40
7.3	REPORTING SUBCATEGORY RELIABILITY.....	41
7.4	INTERRATER CONSISTENCY.....	41
7.5	RELIABILITY OF PERFORMANCE-LEVEL CATEGORIZATION.....	42
7.5.1	Decision Accuracy and Consistency Results.....	43
CHAPTER 8	SCORE REPORTING.....	45
8.1	DECISION RULES.....	46
8.2	STATIC REPORTS.....	46
8.2.1	Student Report.....	46
8.2.2	Summary Reports.....	46
8.3	MONTANA ANALYSIS AND REPORTING SYSTEM.....	48
8.3.1	User Accounts.....	48
8.4	INTERACTIVE REPORTS.....	48
8.4.1	Roster Report.....	49
8.4.2	Performance-Level Summary.....	49
8.4.3	Item Analysis Data.....	50
8.4.4	Longitudinal Data Report.....	50
8.5	INTERPRETIVE MATERIALS AND WORKSHOPS.....	50

8.6	QUALITY ASSURANCE	50
CHAPTER 9	VALIDITY	52
REFERENCES	54
APPENDICES	56
APPENDIX A	ANALYSIS AND REPORTING DECISION RULES	
APPENDIX B	PARTICIPATION RATES	
APPENDIX C	ACCOMMODATION FREQUENCIES	
APPENDIX D	ITEM REVIEW COMMITTEE MEMBERS	
APPENDIX E	ITEM-LEVEL CLASSICAL STATISTICS	
APPENDIX F	ITEM-LEVEL SCORE DISTRIBUTIONS	
APPENDIX G	NUMBER OF ITEMS CLASSIFIED INTO DIFFERENTIAL ITEM FUNCTIONING CATEGORIES	
APPENDIX H	ITEM RESPONSE THEORY CALIBRATION RESULTS	
APPENDIX I	TEST CHARACTERISTIC CURVES AND TEST INFORMATION FUNCTIONS	
APPENDIX J	<i>b</i> -PLOTS	
APPENDIX K	ANALYSES OF EQUATING ITEMS (DELTA AND RESCORE ANALYSES)	
APPENDIX L	SCORE DISTRIBUTIONS	
APPENDIX M	RAW TO SCALED SCORE LOOK-UP TABLES	
APPENDIX N	CLASSICAL RELIABILITY	
APPENDIX O	INTERRATER AGREEMENT	
APPENDIX P	DECISION ACCURACY AND CONSISTENCY RESULTS	
APPENDIX Q	SAMPLE REPORTS	

CHAPTER 1 OVERVIEW OF THE MONTANA CRITERION-REFERENCED TEST

1.1 PURPOSE OF THE ASSESSMENT SYSTEM

The Montana Criterion-Referenced Test (CRT) was developed in accordance with the following federal laws: Title 1 of the Elementary and Secondary Education Act (ESEA) of 1994, P. L. 103–382, and the No Child Left Behind Act (NCLB) of 2001.

In the spring of 2014, Montana students in grades 4, 8, and 10 participated in the Montana CRT in science. This was a reduction from the previous year, in which the CRT covered reading and mathematics in grades 3–8 and 10, in addition to science. The purpose of this assessment is to measure students’ achievement as articulated by Montana’s content standards and grade-level expectations.

All Montana students enrolled in accredited schools are required to participate in the Montana CRT-Science in grades 4, 8, and 10, or the Montana CRT-Alternate. The majority of students use standard administration procedures to participate in the CRT. However, an array of standard accommodations is available to any student, with or without disabilities, when such accommodations are necessary to allow the student to demonstrate his or her skills and competencies. Standard accommodations are not considered to change the constructs being measured and may be provided to students as necessary for any or all of the science portions of the assessment. Students’ tests are scored the same way whether they take the test using standard accommodations or not.

In addition to standard accommodations, other accommodations for the Montana CRT are available to students when specified in their Individual Education Programs (IEP), 504 plans, or limited English proficient (LEP) plans. These other accommodations are referred to as nonstandard accommodations. Because they are considered to alter the constructs being measured, they affect a student’s score on the CRT. When a nonstandard accommodation is used, the student’s score is reported as the lowest possible for that content area (e.g., a scaled score of 200 will fall into the Novice performance level). Nonstandard accommodations may be provided in science, as dictated by the student’s IEP, 504 plan, or LEP plan.

A small percentage of students take the CRT-Alternate to participate in the statewide assessment program. Students with significant cognitive disabilities who are working toward alternate academic achievement standards as documented in their IEPs are eligible to take the CRT-Alternate. Technical characteristics of the CRT-Alternate program are described in a companion technical report.

Montana’s grade- and content-area combination CRT instruments are based on and aligned to Montana’s content standards, benchmarks, and grade-level expectations in science. Montana educators worked with the Montana Office of Public Instruction (OPI) and Measured Progress to develop test items that

assess how well students have met Montana’s grade-level expectations for each content area. In addition, Northwest Regional Educational Laboratory (NWREL) performed an independent alignment study for science in 2007. NWREL’s alignment studies can be found on the OPI’s Web site at www.opi.mt.gov/assessment.

Montana CRT scores are intended to be useful indicators of the extent to which students have mastered material outlined in Montana science content standards, benchmarks, and grade-level expectations. Each student’s Montana CRT score should be used as part of a body of evidence regarding mastery and should not be used in isolation to make high-stakes decisions. Montana CRT scores are more reliable indicators of program success when aggregated to school, system, or state levels, particularly when monitored over the course of several years.

1.2 PURPOSE OF THIS REPORT

This report describes technical aspects of the Montana CRT in an effort to contribute to the accumulation of validity evidence that supports score interpretations of the Montana CRT. Because the interpretations of test scores—not the test itself—are evaluated for validity, this report presents documentation to substantiate intended interpretations (American Educational Research Association [AERA], American Psychological Association & National Council on Measurement in Education, 1999). Subsequent chapters of this report discuss test development and alignment, test administration, scoring, item analyses, equating, reliability and performance levels, and scaled scores and reporting. Each of these topics contributes important information toward establishing the validity of the assessment program. Note, however, that this report does not include certain aspects of a comprehensive validity argument that could also be important to consider when making conclusions about validity. (For instance, additional sources of validity evidence might examine the extent to which Montana CRT scores converge with other measures of the same or similar constructs and diverge from measures of different constructs, or they might examine consequences that arise from scores at the student, school, district, and state levels.)

Historically, some parts of technical reports may have been used by educated laypersons, but the intended audience was experts in psychometrics and educational research. This edition of the Montana CRT technical report attempts to make information more accessible to educated laypersons by providing more thorough descriptions of general categories of information. While making some information more accessible, Measured Progress has also purposely preserved the depth of technical information provided. The reader will find that some discussions and tables continue to require a working knowledge of measurement concepts, such as “reliability” and “validity,” and statistical concepts, such as “correlation” and “central tendency.” To fully understand some of the data presented, the reader will have to possess a basic understanding of advanced topics in measurement and statistics.

CHAPTER 2 ASSESSMENT AND TEST DEVELOPMENT PROCESS

2.1 TEST SPECIFICATIONS

2.1.1 Criterion-Referenced Test

Items on the Montana CRT are developed specifically for Montana and are directly linked to Montana's content standards. These content standards are the basis for the reporting categories developed for each content area and are used to help guide the development of test items. No other content or process is subject to statewide assessment. An item may address part, all, or several of the benchmarks within a standard.

2.1.2 Item Types

Montana educators and students are familiar with the types of items used in the assessment program. The types of items and their functions are described below.

- Multiple-choice items are used to provide breadth of coverage within a content area. Because they require no more than a minute for most students to answer, multiple-choice items make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills.
- Constructed-response items typically require students to use higher-order thinking skills—evaluation, analysis, summarization, etc.—to construct satisfactory responses. Constructed-response items take most students approximately five to 10 minutes to complete. Note that the use of released Montana CRT items to prepare students to respond to constructed-response items is appropriate and encouraged.

2.1.3 Description of Test Design

The Montana CRT is structured using both common and field-test items. Common items are taken by all students in a given grade level. Student scores are based only on common items. In addition, field-test items are divided among the four forms of the test for each grade level. Each student takes only one form of the test and therefore answers a fraction of the field-test items. Field-test items are not identifiable to test takers and have a negligible impact on testing time. Because all students participate in the field test, it provides the sample size (750–1,500 students per item) needed to produce reliable data that can be used to inform item selection for future tests.

2.2 SCIENCE TEST SPECIFICATIONS

2.2.1 Standards

The science specifications are based on Montana’s science content standards.

- **Science Standard 1:** Scientific Investigations—Students, through the inquiry process, demonstrate the ability to design, conduct, evaluate, and communicate results and reasonable conclusions of scientific investigations.
- **Science Standard 2:** Physical Science—Students, through the inquiry process, demonstrate knowledge of properties, forms, changes, and interactions of physical and chemical systems.
- **Science Standard 3:** Life Science—Students, through the inquiry process, demonstrate knowledge of characteristics, structures, and function of living things, the process and diversity of life, and how living organisms interact with each other and their environment.
- **Science Standard 4:** Earth/Space Science—Students, through the inquiry process, demonstrate knowledge of the composition, structures, processes, and interactions of Earth’s systems and other objects in space.
- **Science Standard 5:** Impact on Society—Students, through the inquiry process, understand how scientific knowledge and technological developments impact communities, cultures, and societies.
- **Science Standard 6:** Historical Development—Students understand historical developments in science and technology.

2.2.2 Item Types

The CRT in science includes multiple-choice and constructed-response items. Multiple-choice items require students to select the correct response from four choices, each item taking an average of one minute to answer. Constructed-response items are more involved, requiring five to 10 minutes of response time. Each type of item is worth a specific number of points in the student’s total science score, as shown in Table 2-1.

Table 2-1. 2013–14 Montana CRT: Item Types

<i>Item Type</i>	<i>Possible Score Points</i>
MC	0 or 1
CR	0, 1, 2, 3, or 4

MC = multiple-choice
CR = constructed-response

2.2.3 Test Design

Table 2-2 summarizes the numbers and types of items that were used to compute student scores on the 2013–14 Montana CRT in science. Additionally, each test form had 13 multiple-choice field-test items and one constructed-response field-test item that did not affect student scores.

Table 2-2. 2013–14 Montana CRT: Science Items

Grades	Session 1	Session 2	Session 3	Total	
				MC	CR
4, 8, and 10	17 MC, 1 CR	18 MC	18 MC, 1 CR	53	2

MC = multiple-choice
CR = constructed-response

2.2.4 Blueprints (Distribution of Points across Standards)

Table 2-3 shows the distribution of points and item types across the content standards.

Table 2-3. 2013–14 Montana CRT: Science Specifications/Blueprint—Grades 4, 8, and 10

Montana Standards	Point Distribution by Content Standards	
	Percent	Number
1. Scientific Investigations	23%	14
2. Physical Science	23%	14
3. Life Science	23%	14
4. Earth/Space Science	23%	14
5. Impact on Society		
6. Historical Development	8%	5

The science test design consists of 53 multiple-choice items and two four-point constructed-response items, for 61 total points. In any given year, the two constructed-response items will measure two different standards. From year to year, those standards may change.

2.2.5 Depth of Knowledge

Each item on the Montana CRT in science is assigned a depth of knowledge (DOK) level. The DOK level reflects the complexity of mental processing students use to answer an item. DOK is not synonymous with difficulty. Each of the levels is described below.

- **Level 1 (Recall).** This level requires the recall of information such as a fact, definition, term, or simple procedure. These items require students only to demonstrate a rote response, use a well-known formula, or follow a set procedure.
- **Level 2 (Skill/Concept).** This level requires mental processing beyond that of recalling or reproducing a response. These items require students to make some decisions about how to approach the item.
- **Level 3 (Strategic Thinking).** This level requires reasoning, planning, and using evidence. These items require students to handle more complexity and abstraction than items at the previous two levels.

It is important that the Montana CRT in science measures a range of DOK levels. Table 2-4 shows the percent and point ranges of the three DOK levels used on the Montana CRT in science.

Table 2-4. 2013–14 Montana CRT: DOK Percent and Distribution by Level

<i>DOK Level</i>	<i>Percent Range</i>	<i>Point Range</i>
1	17% to 23%	10 to 14 points
2	56% to 61%	34 to 37 points
3	18% to 23%	11 to 14 points

2.2.6 Use of Calculators and Reference Sheets

Calculators are not used or needed when taking the Montana CRT in science. There are no science reference sheets.

2.3 TEST DEVELOPMENT PROCESS

2.3.1 Item Development

Items used on the Montana CRT are developed and customized specifically for use on the Montana CRT and are consistent with Montana content standards, benchmarks, and grade-level expectations. Measured Progress test developers worked with Montana educators to verify the alignment of items to the appropriate Montana content standards.

The development process combined the expertise of Measured Progress test developers and committees of Montana educators to help ensure items meet the needs of the CRT program. All items used on the common portions of the Montana CRT program were reviewed by a committee of Montana content area experts, as well as a committee of Montana bias experts. Table 2-5 shows the numbers of items developed within each content area for the 2013–14 Montana CRT.

**Table 2-5. 2013–14 Montana CRT: Total Numbers of Items Developed by Content Area—
Grades 3–8 and 10**

MC	CR
75	3

MC = multiple-choice
CR = constructed response

Table 2-6 provides an overview of the item development process for common and field-test items, including the administration of the operational tests.

Table 2-6. 2013–14 Montana CRT: Item Development Process Overview

<i>Development Step</i>	<i>Step Details</i>
Development of items (November 2011 through March 2012)	Measured Progress test developers developed new reading, mathematics, and science items.
Items reviewed for content appropriateness and for bias and sensitivity issues (April 2012)	Committees of Montana educators reviewed the science items.
Edit items (summer 2012)	Montana educators' recommended changes were incorporated into the new items. Measured Progress test developers selected field-test items from the new item pool.
Field-test items (spring 2013)	Embedded field-test items were administered to a sample of students (minimum of 2500 students per item) along with the 2012 operational test.
Statistical review (June 2013)	Montana educators reviewed the field test item statistics and decided which items were acceptable for the item pool.
Item selection (July 2013)	Measured Progress test developers selected common items for the spring 2013 operational CRT tests from the item pool.
Operational test items (March 2014)	Items are part of the common item set and were used to determine student scores. Another embedded field test was also administrated.

2.3.2 Item Reviews at Measured Progress

A science test developer reviewed items for

- item integrity, including content and structure, format, clarity, possible ambiguity, and single correct answer.
- appropriateness and quality of reading selections and graphics.
- appropriateness of scoring guide descriptions and distinctions.
- whether the item is measuring the intended content standard.
- completeness of associated item documentation (e.g. scoring guide, content area codes, key, grade level, DOK, and contract identified).
- appropriateness for the designated grade level.

2.3.3 Item Reviews at State Level

All items were reviewed in Montana. In April 2012, Montana educators from across the state reviewed new items for content appropriateness, alignment to standards, DOK, and grade-level appropriateness.

2.3.4 Bias and Sensitivity Review

Bias review is an essential component of the development process. During the bias review process, items were reviewed by a committee of Montana educators. Items were examined for issues that might offend or dismay students, teachers, or parents. Including such groups in the development of assessment items and materials can avoid many controversial issues, and concerns can be allayed before the test forms are produced.

2.3.5 Reviewing and Refining

Recommended changes from the Item Review and Bias and Sensitivity meetings were incorporated into the items by Measured Progress test developers.

2.3.6 Item Editing

Measured Progress editors then reviewed and edited the items to ensure adherence to sound testing principles and to style guidelines in the *Chicago Manual of Style*, 15th edition. These principles include the stipulations that items

- demonstrate correct grammar, punctuation, usage, and spelling;
- are written in a clear, concise style;
- contain unambiguous explanations that tell students what is required to attain a maximum score;
- are written at a reading level that allows students to demonstrate their knowledge of the subject matter being tested, regardless of reading ability;
- exhibit high technical quality regarding psychometric characteristics;
- have appropriate answer options or score-point descriptors; and
- are free of potentially insensitive content.

2.3.7 Item Selection and Operational Test Assembly

In July 2013, Measured Progress test developers selected common items. In preparation for test construction, test developers and psychometricians at Measured Progress considered the following while selecting sets of items to propose for the common item set to be used on the 2013–14 assessment:

- **Content coverage/match to test design and blueprints.** The test design and blueprints stipulate a specific number of multiple-choice and constructed-response items for each content area. Item selection for the embedded field test was based on the number of items in the existing pool of items that are eligible to be common.

- **Item difficulty and complexity.** Item statistics taken from the data analysis of previously field-tested items were used to ensure similar levels of difficulty and complexity from year to year as well as quality psychometric characteristics.
- **“Cueing” items.** Items were reviewed for any information that might “cue” or provide information that would help answer another item.

2.3.8 Operational Test Draft Review

After the forms were laid out as they would appear in the final test booklets, the forms were again thoroughly reviewed by Measured Progress editors to ensure that the items appeared exactly as intended. Any changes made during test construction were reviewed and approved by the test developer.

2.3.9 Alternative Presentations

Form 1 of each test was translated into Braille by National Braille Press, a subcontractor that specializes in test materials for blind and visually impaired students. In addition, Form 1 for each grade was adapted into a large-print version.

2.4 TEST SESSIONS

The Montana CRT was administered during the spring of 2014 during a four-week period from March 3 to March 25. Science tests were administered in grades 4, 8, and 10. Schools were able to schedule testing sessions at any time during the four-week period, provided they followed the sequence detailed in the scheduling guidelines in the *Test Administrator’s Manual*. Schools were asked to schedule makeup tests for students who were absent from initial test sessions during this testing window.

CHAPTER 3 TEST ADMINISTRATION

3.1 RESPONSIBILITY FOR ADMINISTRATION

As indicated in the *Test Coordinator's Manual*, school system test coordinators, school principals, and/or their designated school test coordinators are responsible for the proper administration of the Montana CRT. This manual was used to ensure the uniformity of administration procedures from school to school.

3.2 ADMINISTRATION PROCEDURES

School test coordinators were instructed to read the *Test Coordinator's Manual* prior to testing and to be familiar with its instructions. The *Test Coordinator's Manual* provides each school with checklists to help coordinators prepare for testing. These checklists outline tasks to be performed before, during, and after test administration. In addition to providing the checklists, the *Test Coordinator's Manual* outlines the nature of the testing materials sent to each school, how to inventory the materials, how to track the materials during administration, and how to return the materials once testing is complete. The *Test Coordinator's Manual* also contains information about including or excluding students.

The *Test Administrator's Manual* includes checklists for administrators to use to prepare themselves, their classrooms, and their students for administration of the test. The *Test Administrator's Manual* contains sections that detail the procedure to be followed for each test session, as well as instructions for preparing the materials prior to giving them to school test coordinators for return to Measured Progress.

The Montana CRT is an untimed assessment; however, guidelines or ranges were provided in the *Test Coordinator's Manual* and the *Test Administrator's Manual* based on the following estimates of the time it takes an average student to respond to each type of item on the test:

- Multiple-choice items—one minute per item
- Constructed-response items—10 minutes per item

The provided guidelines suggested scheduling 45–55 minutes per test session (50–60 minutes for grade 10 students). The guidelines also suggested scheduling a break between each of the three sessions in each content area to prevent test-taker fatigue.

While the guidelines for scheduling were based on the assumption that most students would complete the test within the estimated amount of time, each test administrator was to allow additional time, as necessary, for students to complete the test. If classroom space was not available for this purpose, schools were encouraged to use another space, such as a guidance office. If another space would not be available, the

guidelines recommended scheduling each classroom used for test administration for the maximum possible amount of time.

3.3 PARTICIPATION REQUIREMENTS AND DOCUMENTATION

All students were expected to participate in the Montana CRT; however, the scores of students in the following categories were excluded from the calculation of averages:

- foreign exchange students
- students not enrolled in an accredited Montana school (e.g., home-schooled students)
- students enrolled in a private accredited school
- students enrolled in a private non-accredited school
- students enrolled in a private non-accredited Title 1 school
- students enrolled part-time (less than 180 hours) taking a mathematics or reading course
- first year in U.S. limited English proficiency (LEP) students who were required to participate in the mathematics assessment only
- students who took the CRT using a “nonstandard” accommodation

A summary of this information is shown in Table 3-1, which was published in the *Test Administrator’s Manual* and the *Test Coordinator’s Manual*.

Table 3-1. 2013–14 Montana CRT: Summary of Eligibility for Exclusion from the CRT

<i>Excluded from Averages</i>	<i>MUST Participate</i>	<i>MAY Participate</i>
Foreign exchange students	Yes	
Students not enrolled in an accredited Montana school		Yes
Students enrolled in a private accredited school	Yes	
Students enrolled in a private non-accredited school		Yes
Students enrolled in a private non-accredited Title I school		Yes
Students enrolled part-time (less than 180 hours) taking a mathematics or reading course		Yes
Reading: LEP students in their first year in the United States		Yes

Staff members coded information about any applicable exclusions in the answer booklets after testing was completed. The *Test Coordinator’s Manual* and *Test Administrator’s Manual* provide detailed instructions for coding exclusions and accommodations. In addition, testing exclusions were discussed thoroughly in the pre-administration training audio CD (see Appendix A: Analysis and Reporting Decision Rules).

A summary of participation on the 2013–14 Montana CRT by demographic category for each content area is shown in Appendix B.

3.3.1 Students with Disabilities

All students with special needs participate in the CRT assessment program, by taking either the regular CRT or the CRT-Alternate Assessment if they meet the eligibility criteria.

Form 1 for the grades 4, 8, and 10 tests was enlarged to 18-point font for visually impaired students and was translated into Braille by National Braille Press, a subcontractor that specializes in test materials for blind students. Students with special needs and LEP students are often given these test accommodations.

3.4 ADMINISTRATOR TRAINING

The OPI hosted a test-administration workshop in Helena, Montana, on January 8–10, 2014. The workshop was well attended, but attendance of system and school test coordinators was not mandatory. OPI and Measured Progress staff members hosted sessions that covered test accommodations, student information system (AIM) updates, CRT materials and administration, CRT-Alternate materials and administration, online reporting, and test security.

In addition to the workshop and the distribution of the *Test Coordinator's Manual* and *Test Administrator's Manual*, the OPI and Measured Progress produced the PowerPoint presentation “Spring 2014: CRT and CRT-Alt Overview” The training materials and the PowerPoint presentation were posted on the OPI's Web site: <http://www.opi.mt.gov>. The PowerPoint presentation provided the training information for system and school test coordinators who were unable to attend the administration workshop. The PowerPoint presentation also served as a useful tool for training both system and school personnel.

3.5 DOCUMENTATION OF ACCOMMODATIONS

The *2014 CRT Accommodations Manual* and the accommodations training PowerPoint presentation “Guidelines and Procedures for CRT Accommodations” were produced by the OPI and posted online at <http://opi.mt.gov/Curriculum>. General instructions regarding accommodation usage and a list of available accommodations were included in the *Test Coordinator's Manual*.

Standard accommodations were available to all students on the basis of individual needs regardless of disability status. Decisions regarding standard accommodations were made by the student's educational team on an individual basis and were consistent with either previous accommodation decisions for the student or current educational needs. Accommodations usage was required to be consistent with those used during the student's regular classroom instruction and assessment for at least three months prior to testing.

Nonstandard accommodations were offered to students with disabilities only if the accommodations were specified in the student’s Individual Educational Program (IEP). If a student was assessed with a nonstandard accommodation, the student was considered a nonparticipant in calculations of the participation rate for average yearly progress (AYP) purposes. In addition to the student being considered a nonparticipant, the student’s score from the assessment was not included in calculations of the proficiency rate for AYP.

Table 3-2 shows the number of students at each grade who were tested with and without accommodations. In addition, the numbers of students who were tested with accommodations are presented by accommodation type in Appendix C.

**Table 3-2. 2013–14 MontCAS: Numbers of Students Tested With and Without Accommodations—
Science**

Grade	Number of Students Tested	
	With Accommodations	Without Accommodations
4	1,068	9,777
8	664	9,679
10	526	9,626

3.6 TEST SECURITY AND ADMINISTRATION IRREGULARITIES

Test coordinators and administrators are prohibited from disclosing the contents of CRT assessments. Under no circumstances are test booklets or marked answer booklets circulated among faculty, administrators, or other persons.

All system test coordinators and school principals received the *OPI Guidelines and Procedures for Test Security*. This OPI publication was made available to system superintendents, principals, and test administrators to outline the reporting procedures for security and administration violations. All concerns about breaches of test security or noncompliance with test administration procedures were to be reported immediately to the principal, system test coordinator, and state assessment director.

3.7 TEST ADMINISTRATION WINDOW

The Montana CRT was administered during the spring of 2014 during a four-week period from March 3 to March 25. Science tests were administered in grades 4, 8, and 10. Schools were able to schedule testing sessions at any time during the four-week period, provided they followed the sequence detailed in the scheduling guidelines in the *Test Administrator’s Manual*. Schools were asked to schedule makeup tests for students who were absent from initial test sessions during this testing window.

3.8 SERVICE CENTER

To address testing concerns, Measured Progress established a help desk dedicated to the Montana CRT. Service Center support is an essential element to the successful administration of large-scale assessments. It provides a central location that individuals in the field can call via a toll-free number to request assistance, report problems, or ask specific questions.

The Measured Progress help desk provided support during all phases of the testing window. It was staffed at varying levels, based on need and the volume of calls received, from 8:00 a.m. to 4:00 p.m. MST. At a minimum, the help desk consisted of a product support specialist responsible for receiving, responding to, and tracking calls and e-mails, and routing issues to the appropriate person(s) for resolution. In addition, the program manager and/or program assistant addressed communications that required a higher level of program support.

During the period between February 6, 2014, when the testing materials were delivered to schools, and April 11, 2014, when the materials were returned to Measured Progress, the Service Center received approximately 35 calls. The majority of these calls were to order additional materials for students who enrolled after materials were shipped and to arrange for UPS to pick up the materials after testing. The service center staff also responded to administration questions and referred policy questions regarding test security or accommodations usage to the OPI.

CHAPTER 4 SCORING

Accurate and timely scoring of constructed-response, short-answer, and multiple-choice items is an important process in any successful assessment program. This chapter defines the scope and processes of Measured Progress’s scoring services for the 2013–14 Montana CRT.

4.1 MACHINE-SCORED ITEMS

Preceding the arrival of the Montana CRT answer booklets, Measured Progress prepared customized scanning programs to enable selective reading of all the scannable materials that included student identification and demographics and to electronically format the scanned information.

Once the student answer booklets were received from each Montana school following test administration, Measured Progress optically scanned each page from every answer booklet to create digital images of the entire document. Every page was bar-coded so that the scores applied to each item could be linked to the correct student, school, and district. Student responses were then imported into iScore, Measured Progress’s proprietary image-based scoring system, for secure processing and scoring. By using this system, Measured Progress was able to increase reliability and productivity, as well as monitor and maintain quality control.

Student multiple-choice response data were machine-scored at the same time that student constructed-response and short-answer items were scanned into iScore for hand-scoring. Multiple-choice items were compared to scoring keys via item-analysis software. Correct multiple-choice answers were assigned a score of one point, and incorrect answers were assigned zero points. Student multiple-choice responses consisting of multiple marks and blank responses were also assigned zero points.

Student responses that could not physically be scanned (e.g., documents damaged during administration or shipment) were physically reviewed and scored on an individual basis by trained, qualified staff. These scores were linked to the student’s demographic data and merged with the student’s scoring file by Measured Progress’s Data and Reporting Services department.

Table 4.1 shows the number of responses scanned and scored for each grade in science.

Table 4-1. 2013–14 Montana CRT: Number of Responses Scanned and Scored—Science

<i>Grade</i>	<i>Number of Responses Scanned and Scored</i>
4	22,035
8	21,134
10	21,194

4.2 PERSON-SCORED ITEMS

Scanned images of open-response items were processed and organized into item-specific groups in preparation for person-scoring by iScore. iScore’s secure, Web-based application provided qualified staff, including readers and their leadership, password-protected access for reading and scoring electronic student responses at one or multiple scoring sites without compromising confidentiality. The digital image clip information of constructed-response and short-answer responses allowed iScore to replicate student responses just as they appeared on the originals and to display the replicated responses on individual monitors for person-scoring. In addition, the processes of item benchmarking, reader training, scoring, editing/cleanup, and reporting were all accomplished electronically and without further reference to the originals.

Organized by iScore in this way, qualified readers were able to view only one response from a single item at a time. Because item responses were tracked and distributed among groups of readers by iScore, each response in an individual student’s response booklet could be assigned to and scored by a different reader. This maximization of the number of readers per student response booklet effectively minimized bias errors caused by reader sampling.

Leadership staff, on the other hand, had constant, albeit view-only, access to all of the imaged responses from a student’s booklet for whenever necessary. The actual test booklets and answer documents were also available to the content area chief reader and the iScore operational manager (see section on “Scoring Location and Staff”).

To ensure the security of constructed-response and short-answer items and responses scored, all scoring activities in iScore were performed “blind,” i.e., without student names, district, and/or school information visible or able to be associated with responses or raw scores. During scoring, iScore distributed images of student responses to the computer monitors of readers located at one of Measured Progress’s scoring facilities. When iScore sent an image of a student response to an individual reader’s computer monitor, the reader evaluated the response and recorded the score via keypad or mouse entry. Once the score was entered, a new response appeared immediately on the screen.

Although iScore is based on conventional, best-practice scoring procedures, it also offers the following benefits:

- It provides leadership staff with real-time information about group- and individual-level performance, including scoring accuracy and consistency, as well as overall process monitoring and reporting.
- It ensures the randomized distribution of student responses among readers during scoring and automatically assigns student responses to one or more scorers for interrater agreement monitoring.
- It permits password-only access limited to those solely in the employ of Measured Progress and working within a qualified scoring or scoring-management capacity.

- It maintains student anonymity and confidentiality by masking student biographical information from viewers.
- It offers immediate access to samples of student responses and scores for reporting and analysis.
- It offers early access to subsets of data for tasks such as standard setting.
- It reduces material handling, which saves time and labor while enhancing the security of materials.

The iScore database, its control operation, and its administrative offices are all based in Dover, New Hampshire. The iScore system monitored accuracy, reliability, and consistency across all Measured Progress scoring facilities. To ensure that scoring information and updates were equally shared and implemented across all scoring facilities, constant communication and coordination was accomplished daily via e-mail, telephone, fax, and secure, Web-based networks.

4.2.1 Scoring Location and Staff

Scoring Location

Scoring of the 2013–14 Montana CRT in science took place at Measured Progress’s scoring facilities located in Menands, New York.

Scoring Staff

Staffing for the 2013–14 Montana CRT implemented low scoring leadership-to-reader ratios and was composed of the following Measured Progress staff members:

- A scoring project manager, who oversaw the overall contract from a scoring perspective and acted as a liaison with contract management staff, data analysis staff, and the client while managing the content area experts (chief readers, quality assurance coordinators, etc.).
- A chief reader, who prepared benchmarking/training materials and led the review and client approval of materials, working closely with Measured Progress test developers and Montana educators. The Chief reader trained, qualified, and monitored readers during the scoring process; supervised quality assurance coordinators, senior readers (SR), and readers; and monitored scoring accuracy and consistency. The ratio of chief readers to the scoring project manager was one to one.
- Quality assurance coordinators (QAC), who managed the training and benchmarking of items for each grade within the Montana CRT. QACs trained, qualified, and monitored readers during the scoring process, supervised SRs and readers, and monitored scoring accuracy and consistency.
- SRs, who supervised readers during the scoring process and monitored scoring accuracy and consistency while managing quality control measures via iScore. During scoring, the ratio of SRs to QACs was one to one.

- Readers, who were qualified temporary staff members performing the bulk of scoring work, evaluated and scored student responses according to the Montana CRT guidelines provided for each grade level and content area scored. Readers received the same orientation and training as direct hires. The ratio of readers to SRs varied by grade but did not exceed 11 to one.

4.2.2 Reader Recruitment and Qualifications

In preparation for scoring the 2013–14 Montana CRT, Measured Progress actively sought and recruited readers to represent a diverse spectrum of educational, professional, and ethnic populations. The customary cross section of readers employed included business professionals, scientists, graduate school students, and both current and retired educators.

Although the employment of readers holding a four-year college degree or higher was preferred, all readers were required to have successfully completed a minimum of at least two years of college and to have demonstrated knowledge of the content area they scored. All readers were required to submit documentation (i.e., college transcript and/or resume) of their qualifications.

For training and qualification, readers were placed at grade levels that matched their areas of experience and expertise. Reader demographic information (gender, education, ethnic background, etc.) was electronically documented for reporting. All readers were subject to stringent nondisclosure requirements and supervision and were required to sign a nondisclosure/confidentiality agreement. Table 4-3 summarizes the educational credentials of the 2013–14 Montana CRT readers and QACs.

Table 4-2. 2013–14 Montana CRT: Education Credentials of Readers and QACs

	<i>Description</i>	<i>Menands, NY</i>	<i>Total</i>	<i>Percent</i>
Readers	Less than 48 college credits	0	0	0.00
	48+ college credits	0	0	0.00
	Associate’s degree	1	1	2.60
	Bachelor’s degree	25	25	64.10
	Master’s degree	12	12	30.70
	Doctorate	1	1	2.60
	Total	39	39	100
QACs	Less than 48 college credits	0	0	0
	48+ college credits	0	0	0
	Associate’s degree	0	0	0
	Bachelor’s degree	7	7	87.50
	Master’s degree	1	1	12.50
	Doctorate	0	0	0
	Total	8	8	100

4.2.3 Methodology for Scoring Polytomous Items

Possible Score Points

The ranges of possible score points for the different polytomous items found on the 2013–14 Montana CRT are blank (B), 0, 1; and B, 0, 1, 2, 3, 4.

Condition Codes

When numerical score-point parameters did not apply to a student response, readers had the option of designating one of the following options:

- blank response (empty entry without an attempt at responding to the question)
- unreadable response (illegible response or too faint to accurately interpret)
- wrong location (a relevant response entered into the space reserved for a different item)
- non-English response (a response written entirely in a language other than English)

Responses designated unreadable and wrong location were resolved by consulting the original test booklet and/or by identifying the correct location.

4.2.4 Reader Training

For each item scored in the 2013–14 Montana CRT, Measured Progress readers were required to demonstrate their scoring ability by participating in training sessions specific to each item scheduled to be scored. The scoring project began with an introduction of the onsite scoring staff and an overview of the Montana CRT program’s purpose and goals (including discussion about document security, student confidentiality, the proprietary nature of testing materials, scoring materials, and iScore procedures).

Actual training began with groups of readers organized into grade-, and item-specific group assignments. Each reader was provided with a personal hard copy of item-specific training materials distributed at the beginning of each work session and had to account for these materials during secure collection at the end of each work session. During training, readers were strongly encouraged to take notes and highlight their hard copies of the training materials.

For each item trained, the QAC assigned to the item commenced reader training by reviewing and discussing the prompt and item-specific scoring guide. The training QAC demonstrated the process of applying the item’s scoring guide and score-point descriptors to the exemplars found in the subsequent anchor and training sets before attempting to demonstrate scoring accuracy in the qualifying set.

Anchor Set

An anchor set is a set of responses approved by the respective content-area specialists representing the OPI for reading, mathematics, or science. Each anchor set contained at least one OPI-approved sample response per score point considered to be a mid-range exemplar. Responses in the anchor sets were typical,

rather than unusual or uncommon; solid, rather than controversial or borderline; and true, meaning that their scores could not be changed except by the OPI and Measured Progress test developers.

Training QACs facilitated group discussion of anchor set responses in relation to the scoring guide and score-point descriptors to help the readers internalize the typical characteristics of score points. The anchor set served as a reference for readers as they went on to score sample responses in the training set that followed.

Training Set

Next, readers practiced applying the scoring guide and anchor set to responses in the training set. The training set typically included six to 10 student responses designed to help establish both the full score-point range and the variation of possible responses within each score point. The training set often included unusual responses that were less clear or solid (e.g., briefer than normal, employing atypical approaches, simultaneously containing very low and very high attributes, and written in ways difficult to decipher).

Responses in the training set were presented to readers without scores and in a randomized score-point order. Once readers had independently read and applied their score to a training set response, the training QAC would discuss how the response was actually scored. The QAC then responded to reader questions and/or comments while pointing out particular scoring issues at hand (e.g., the borderline between two score points). Throughout each item training, the QAC continually routed reader discussion of score points back to the anchor set and scoring guide. After the training set had been completed, readers were required to use qualifying sets that were assembled from constructed-response items to demonstrate their scoring accuracy.

Qualifying Set

Following participation in each item training session, readers were administered a qualification set of committee-reviewed responses (CRR) assigned to each item in the reader's content area. Each qualifying set was composed of 10 preselected, previously scored responses chosen as clear illustrations of score-point examples that would measure which readers had adequately internalized item training before those readers were able to score live student responses. These CRRs were selected by scoring leadership and randomly distributed to each reader via iScore during qualification.

In order to qualify on a qualification set, readers were required to demonstrate a scoring accuracy level of at least 80% exact agreement (i.e., exactly match scores on at least eight of the 10 CRRs) and at least 90% exact-or-adjacent agreement (i.e., exactly match or be within one score point on nine or 10 of the 10 CRRs). In other words, readers were allowed one discrepant score (i.e., one score out of the 10 CRRs that was more than one score point from the CRR score) provided they had at least eight exact scores.

Once a group of readers successfully qualified on a particular item, responses for that item in iScore were randomly assigned and presented to them on their computer monitors, one response at a time. Readers

unable to qualify on the first qualification set received QAC retraining (see section on “Retraining”) and a subsequent opportunity at qualification on a second qualification set. Readers unable to qualify on the second qualification set were not eligible to score that item.

(Note: In order to be eligible to score short-answer mathematics items in grades 3–8 and 10, readers were required to qualify on at least one mathematics constructed-response item for that grade.)

Retraining

Readers unable to qualify on the first qualification set received QAC retraining by reviewing their performance in relation to the item training materials. The QAC responded to reader questions and routed discussion of score points back to the anchor set and scoring guide. Readers were then allowed the opportunity at qualification on a second qualification set. Readers unable to qualify on the second qualification set were not eligible to score that item. Table 4-4 depicts the accuracy and qualification percentages of the reader applicants.

Table 4-3. 2013–14 Montana CRT: Scoring Accuracy and Qualification Statistics

Content Area	Grade	Item	Average Percent Exact Agreement		Readers		
			Embedded CR Sets*	Double-Blind Scoring*	Taking Qualification Sets	Successfully Qualifying	Percent Qualifying
Science	4	23	89.9	72.4	12	12	100
		69	77.7	77.2	12	12	100
	8	23	89	72.9	14	14	100
		69	86.1	75.6	13	13	100
	10	23	95.7	92	7	7	100
		69	76.8	85.8	13	13	100

* Embedded and double-blind accuracy rates are calculated by excluding scoring by leadership (SR/QAC or chief reader), who score a minimal number of responses as part of clean up and do not score enough to trigger these quality control measures.

4.2.5 Leadership Training

A core group of scoring leadership staff, including QACs and SRs, guided and monitored readers throughout the scoring process. Because quality control by QACs and SRs moderated the scoring process and maintained the integrity of scores, the individuals chosen to fill those positions were selected for their accuracy and consistency. The training QACs assigned to train readers were also selected for their ability to instruct, as well as for their content-area specialization.

The purpose of leadership training was to ensure that QACs provided thorough and consistent training and feedback to readers. Chief readers were able to discuss item details and score-point rationale within training materials in order to prepare scoring leadership for reader questions. Chief readers reviewed items with QACs, who in turn trained and reviewed items with their SRs and readers. During actual item

scoring, QACs trained and supervised readers and monitored SR accuracy and consistency. The SRs, in turn, supervised their own group of readers and monitored reader accuracy and consistency.

4.2.6 Monitoring of Scoring Quality Control

iScore was preprogrammed to monitor individual reader accuracy and scoring consistency among readers on a constant basis. iScore's use of multiple monitoring techniques was critical to the process of live scoring, allowing readers who met or exceeded accuracy standards to continue scoring. Reader accuracy and consistency was measured in iScore throughout the scoring process using the following methods and tools:

- embedded CRRs
- read-behind scoring
- double-blind scoring
- reader arbitration

Embedded CRRs

Embedded CRRs are preselected, previously scored responses used to ensure that readers had adequately internalized item training and remained calibrated to the scoring standard during live scoring. Prior to scoring, scoring leadership selected embedded CRRs for each item and loaded the examples into. Each example represented images of actual student work and appeared no different from live student responses. During the first day of live scoring of each item, iScore randomly distributed 30 embedded CRRs to each reader. Embedded CRRs were employed for all constructed-response items and enabled scoring leadership to monitor reader accuracy and consistency as gauged by the known scores of the embedded CRRs.

Readers with a disproportionate number of adjacent and/or discrepant scores in embedded CRRs were subject to increased monitoring, additional read-behinds, consultation by scoring leadership, and/or retraining by the QAC. Following these measures, it was at the discretion of scoring leadership whether or when the reader could resume scoring. If the individual was allowed to resume scoring, scoring leadership carefully monitored these readers by increasing the number of read-behinds.

Read-Behind Procedures

Read-behind scoring refers to scoring leadership (typically a SR) scoring a response that was recently scored by a reader. The gain was an immediate, real-time snapshot of each reader's accuracy and consistency during scoring. SRs were required to perform read-behinds on each reader throughout each day and at any point during scoring. This practice was applied to all open-ended item types and performed on all readers as required.

Once called up in iScore by the SR, read-behind responses were selected by iScore and placed into the SR's read-behind queue. Readers were aware of neither iScore's selection nor which of their scored

responses was to be reviewed by their SR. Likewise, SRs were not aware of the reader’s score when iScore presented each read-behind response for their own review and eventual score. The SR then applied his or her own score to the response before the reader’s score was made viewable in iScore. This SR review and comparison of the two scores created the score-of-record determination (i.e., the reported score) as follows:

- If the reader and SR applied the same score (exact agreement), no action was necessary; the reader’s score became the score of record.
- If the reader and SR scores differed by one point (adjacent), the SR’s score became the score of record, thereby overriding the reader’s score.
- If the reader and SR scores differed by more than one point (discrepant), the SR’s score became the score of record, thereby overriding the reader’s score.

Readers with a disproportionate number of adjacent and/or discrepant scores in read-behinds were subject to increased monitoring, additional read-behinds, consultation by scoring leadership, and/or retraining by the QAC. Following these measures, it was at the discretion of scoring leadership whether or when the reader could resume scoring. If the individual was allowed to resume scoring, scoring leadership carefully monitored these readers by increasing the number of read-behinds. Table 4-5 outlines the resolution of reader scores using the read-behind procedure.

Table 4-4. 2013–14 Montana CRT: Examples of Read-Behind Scoring Resolution

<i>Reader Score</i>	<i>QAC/SR Score</i>	<i>Score of Record</i>
4	4	4
4	3	3*
4	2	2*

* QAC/SR’s score

Double-Blind Scoring

Scoring procedures for both constructed-response and short-answer item types included both single-scoring and double-scoring. Single-scored responses were scored by one reader. Double-scored responses were scored “blindly” by two different readers, unaware of the other’s score. These double-blind scores were monitored for interrater-agreement accuracy and scoring consistency. A default minimum setting of 2% of all open-ended item types was double-blind scored. In addition, responses marked blank were automatically routed for double-blind scoring. Table 4-6 indicates the frequency for which open-ended item responses from each content area were double-blind scored.

Table 4-5. 2013–14 Montana CRT: Frequency of Double-Blind Scoring

<i>Grade</i>	<i>Content Area</i>	<i>Responses Double-Blind Scored</i>
4, 8, 10	Science	2% minimum

All	Blank responses	100%
-----	-----------------	------

Reader Arbitration

When double-blind scores applied by two readers on a single response differed by more than one point (a discrepancy), iScore placed the response into an arbitration queue for review and rescoring by the SR. Readers were aware neither of the discrepancy arbitration nor which of their scored responses was to be arbitrated. Likewise, the SR was not aware of either readers' scores when iScore presented the response for review. It was only after the SR had applied his or her own score to the response that the readers' scores were made viewable. This SR review and rescoring of the response became the score of record, thereby overriding the readers' scores.

Readers with a disproportionate number of adjacent and/or discrepant scores in double-blind scoring were subject to increased monitoring, additional read-behinds, consultation by scoring leadership, and/or retraining by the QAC. Following these measures, it was at the discretion of scoring leadership whether or when the reader could resume scoring. If the individual was allowed to resume scoring, scoring leadership carefully monitored these readers by increasing the number of read-behinds. Table 4-7 displays the final summary statistics for double-blind scoring.

Table 4-6. 2013–14 Montana CRT: Double Blind Summary Statistics

<i>Content Area</i>	<i>Grade</i>	<i>Number Scored</i>	<i>Responses</i>	
			<i>Double-Blind Scored Number</i>	<i>Percent</i>
Science	4	22,035	711	3.2%
	8	21,134	807	3.8%
	10	21,194	1445	6.8%

In the case that the individual was not allowed to resume scoring, the content area chief reader had the right to remove (“void”) all of that reader’s scores applied to the item from that day’s work totals. Voided responses in iScore were returned to the response queue and rescored by readers able to maintain the scoring accuracy standard. Table 4-8 summarizes the statistics relevant to voided or blocked readers.

Table 4-7. 2013–14 Montana CRT: Voided or Blocked Reader Statistics

<i>Content Area</i>	<i>Grade</i>	<i>Item</i>	<i>Number of Readers</i>	
			<i>With Voided Scores</i>	<i>NOT Allowed to Continue Scoring</i>
Science	4	23	0	0
		69	0	0
	8	23	0	0
		69	3	0
	10	23	0	0
		69	2	0

CHAPTER 5 CLASSICAL ITEM ANALYSIS

As noted in Brown (1983), “A test is only as good as the items it contains.” A complete evaluation of a test’s quality must include an evaluation of each item. Both *Standards for Educational and Psychological Testing* (AERA et al., 1999) and *Code of Fair Testing Practices in Education* (2004) include standards for identifying quality items. Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. Items should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. In addition, items must not unfairly disadvantage students in particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that Montana CRT items meet these standards. Qualitative analyses are described in earlier chapters of this report; this chapter focuses on quantitative evaluations. Statistical evaluations are presented in four parts:

- difficulty indices
- item-test correlations
- differential item functioning (DIF) statistics
- dimensionality analyses.

The item analyses presented here are based on the statewide administration of the Montana CRT in spring 2013. Note that the information presented in this chapter is based on the items common to all forms, since those are the items on which student scores are calculated. (Item analyses are also performed for field-test items, and the statistics are then used during the item-review and form-assembly processes for future administrations.)

5.1 CLASSICAL DIFFICULTY AND DISCRIMINATION INDICES

All multiple-choice, constructed-response, and short-answer items are evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing it by the maximum possible score for the item. Multiple-choice and short-answer items are scored dichotomously (correct vs. incorrect), so for these items the difficulty index is simply the proportion of students who correctly answered the item. Constructed-response items are scored polytomously, meaning that a student can achieve a score of 0, 1, 2, 3, or 4. By computing the difficulty index as the average proportion of points achieved, the indices for the different item types are placed on a similar scale, ranging from 0.0 to 1.0, regardless of the item type. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an easiness index, because larger values indicate easier items. An index of 0.0

indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student abilities, but they do indicate knowledge or skills that have been mastered by most students. Similarly, items that are correctly answered by very few students provide little information about differences in student abilities, but may indicate knowledge or skills that have not yet been mastered by most students. In general, to provide the best measurement, difficulty indices should range from near-chance performance (0.25 for four-option multiple-choice items, or essentially zero for constructed-response or short-answer items) to 0.90, with the majority of items generally falling between around 0.4 and 0.7. However, on a standards-referenced assessment such as the Montana CRT, it may be appropriate to include some items with very low or very high item-difficulty values to ensure sufficient content coverage.

A desirable characteristic of an item is for higher-ability students to perform better on the item than lower-ability students. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of the item. Within classical test theory, the item-test correlation is referred to as the item's discrimination, because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For constructed-response items, the item-discrimination index used was the Pearson product-moment correlation; for dichotomous items (multiple-choice and short-answer), the corresponding statistic is commonly referred to as a point-biserial correlation. The theoretical range of these statistics is -1.0 to 1.0 , with a typical observed range from 0.2 to 0.6 .

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency.

A summary of the item-difficulty and item-discrimination statistics for science is presented in Table 5-1. Note that the statistics are presented for all items, as well as by item type (multiple-choice and open-response, which includes both constructed-response and short-answer items). The mean difficulty and discrimination values shown in the table are within generally acceptable and expected ranges.

Table 5-1. 2013–14 MontCAS: Summary of Item Difficulty and Discrimination Statistics—Science

Grade	Item Type	Number of Items	p-value		Discrimination	
			Mean	Standard Deviation	Mean	Standard Deviation
4	ALL	55	0.65	0.15	0.33	0.07
	MC	53	0.66	0.14	0.32	0.07
	OR	2	0.34	0.11	0.43	0.11
8	ALL	55	0.61	0.13	0.32	0.09
	MC	53	0.62	0.12	0.31	0.08
	OR	2	0.37	0.11	0.48	0.07
10	ALL	55	0.6	0.17	0.33	0.11
	MC	53	0.61	0.16	0.32	0.1
	OR	2	0.32	0.01	0.57	0.06

MC = multiple-choice
OR = open-response

5.2 DIFFERENTIAL ITEM FUNCTIONING

A comparison of indices across grade levels is complicated, because these indices are population dependent. Direct comparisons would require that either the items or students were common across groups. Since that is not the case, it cannot be determined whether differences in performance across grade levels are due to differences in student abilities, differences in item difficulties, or both. With this caveat in mind, it appears that for science, students in higher grades found their items more difficult than students in lower grades found theirs.

Comparing the difficulty indices of multiple-choice and constructed-response or short-answer items is inappropriate, because multiple-choice items can be answered correctly by guessing. Thus, it is not surprising that the difficulty indices for multiple-choice items tend to be higher (indicating that students performed better on these items) than the difficulty indices for constructed-response items. Similarly, discrimination indices for the four-point constructed-response items were larger than those for the dichotomous items due to the greater variability of the former (i.e., the partial credit these items allow) and the tendency for correlation coefficients to be higher given greater variances of the correlates.

In addition to the item-difficulty and discrimination summaries presented above, item-level classical statistics and item-level score distributions were also calculated. Item-level classical statistics are provided in Appendix E; item-difficulty and discrimination values are presented for each item. The item-difficulty and discrimination indices are within generally acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. There were a small number of items with near-zero discrimination indices, but none were negative. While it is not inappropriate to include items with low discrimination values or with very high or very low item-difficulty values to ensure that content is appropriately covered, there were very few such cases on the Montana CRT. Item-level score-

point distributions are provided for constructed-response items in Appendix F; for each item, the percentage of students who received each score point is presented.

5.3 DIMENSIONALITY ANALYSIS

The DIF analyses of the previous section were performed to identify items which showed evidence of differences in performance between pairs of subgroups beyond that which would be expected based on the primary construct that underlies total test score (also known as the “primary dimension;” for example, general achievement in math). When items are flagged for DIF, statistical evidence points to their measuring an additional dimension(s) to the primary dimension.

Because tests are constructed with multiple content area subcategories, and their associated knowledge and skills, the potential exists for a large number of dimensions being invoked beyond the common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional IRT models that are used for calibrating, linking, scaling, and equating the 2013-14 MontCAS test forms. As noted in the previous section, a statistically significant DIF result does not automatically imply that an item is measuring an *irrelevant* construct or dimension. An item could be flagged for DIF because it measures one of the construct-*relevant* dimensions of a subcategory’s knowledge and skills.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Findings from dimensionality analyses performed on the 2013-14 MontCAS common items for science are reported below. (Note: only common items were analyzed since they are used for score reporting.)

The dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Both of these methods use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on expected total score for the rest of the test, and the average conditional covariance is obtained by averaging over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected total test scores. Non-zero conditional covariances are essentially violations of the principle of local independence, and local *dependence* implies multidimensionality. Thus, non-random patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. The data are first divided into a training sample and a cross-validation sample.

Then an exploratory analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items displays local dependence, conditioning on total score on the non-clustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. As with DIMTEST, the data are first divided into a training sample and a cross-validation sample. The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample are used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances are summed, from this sum the between-cluster conditional covariances are subtracted, this difference is divided by the total number of item pairs, and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality), values of 0.2 to 0.4 weak to moderate multidimensionality; values of 0.4 to 1.0 moderate to strong multidimensionality, and values greater than 1.0 very strong multidimensionality.

DIMTEST and DETECT were applied to the 2013-14 MontCAS science tests. The data for each grade level (4, 8, and 10) were split into a training sample and a cross-validation sample. Every grade level had at least 10,300 student examinees, so every training sample and cross-validation sample had at least 5,150 students. DIMTEST was then applied to every grade level. DETECT was applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

Because of the large sample sizes of the Montana tests, DIMTEST would be sensitive even to quite small violations of unidimensionality, and the null hypothesis was rejected at a significance level of 0.01 for every dataset. The rejection of the null hypothesis of unidimensionality for every test was not surprising because strict unidimensionality is an idealization that almost never holds exactly for a given dataset. Thus, it was important to use DETECT to estimate the effect size of the violations of local independence found by DIMTEST. Table 8-11 displays the multidimensional effect size estimates from DETECT.

Table 5-2. 2013–14 MontCAS: Multidimensionality Effect Sizes by Content Area and Grade

<i>Content Area</i>	<i>Grade</i>	<i>Multidimensionality Effect Size</i>	
		<i>2013–14</i>	<i>2012–13</i>
Science	4	0.09	0.13
	8	0.12	0.13
	10	0.12	0.13
	Average	0.11	0.13

All the DETECT values for 2013-14 indicated very weak multidimensionality. Also shown in Table 8-11 are the values reported in last year's dimensionality analyses. The DETECT indices for each grade are seen to be similar between the two years. In particular, both sets of values indicate very weak multidimensionality for all the tests. We also investigated how DETECT divided the tests into clusters to see if there were any discernable patterns with respect to item type – that is, multiple choice (MC) and constructed response (CR). Because there were only two CR items at each grade level, it was difficult to judge whether the clusters produce a significant separation of the MC and CR items. The strongest separations occurred with grades 4 and 8. In grade 4, the investigation of the sign patterns in the conditional covariance matrix indicated the CR items formed a single cluster with a positive conditional covariance between the two CR items but with no systematic positive conditional covariance with the MC items. In grade 8, inspection of the conditional covariance sign pattern matrix revealed that each CR item seemed to form its own cluster, with only a few MC items having positive conditional covariances with them. It is interesting to note that similar results also occurred in 12-13 in that science grade 4 and science grade 8 also displayed evidence of MC-CR separation in last year's analysis. This is the third year in a row that grade 8 science has displayed MC-CR separation, while grade 4 science has done so for four out of the past six years. Science grade 10 showed no evidence of systematic MC-CR separation, as also was observed for the 12-13 test. A more thorough investigation employing experts in the substantive content of the test forms may result in identification of clusters related to the skills and knowledge areas measured by the items. In any case the violations of local independence from all such effects, as evidenced by the DETECT effect sizes, were very small and do not warrant any changes in test design or scoring.

CHAPTER 6 ITEM RESPONSE THEORY SCALING AND EQUATING

This chapter describes the procedures used to calibrate, equate, and scale the Montana CRT. During the course of these psychometric analyses, a number of quality-control procedures and checks on the processes were implemented. These procedures included evaluation of the calibration processes (e.g., checking the number of Newton cycles required for convergence for reasonableness, checking item parameters and their standard errors for reasonableness, or examining test characteristic curves [TCC] and test information functions [TIF] for reasonableness), evaluation of model fit, evaluation of equating items (e.g., delta analyses, rescore analyses, examination of *b*-plots for reasonableness) and evaluation of the scaling results (e.g., parallel processing by the Psychometrics and Research and Data Analysis departments, comparing lookup tables to the previous year’s). An equating report, which provided complete documentation of the quality-control procedures and results, was reviewed by the Montana Department of Education and approved prior to production of student reports (Measured Progress Department of Psychometrics and Research, *2013–14 MontCAS Criterion-Referenced Test Equating Report*, unpublished manuscript).

Table 6-1 lists items that required intervention either during item calibration or as a result of the evaluations of the equating items. For each flagged item, the table shows the reason it was flagged (e.g., the item was flagged as a result of the delta analyses) and what action was taken. The number of items identified for evaluation was typical across grades and content areas. Descriptions of the evaluations and results are included in the Item Response Theory Results and Equating Results sections below.

Table 6-1. 2013–14 MontCAS: Items that Required Intervention During IRT Calibration and Equating—Science

<i>Grade</i>	<i>IABS</i>	<i>Reasons</i>	<i>Action</i>
4	120019	item fit	retained for equating
	166239	item fit	retained for equating
	209656	item fit	retained for equating
8	210221	b/b analysis	removed from equating
		delta analysis	removed from equating
	75906	b/b analysis delta analysis	removed from equating removed from equating
10	119654	item fit	retained for equating
	134497	item fit	retained for equating
	75735	item fit	retained for equating

6.1 ITEM RESPONSE THEORY

All Montana CRT items were calibrated using item response theory (IRT). IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as

theta (θ), and the probability (p) of getting a dichotomous item correct or of getting a particular score on a polytomous item (Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985). In IRT, it is assumed that all items are independent measures of the same construct (i.e., of the same θ). Another way to think of θ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between θ and p (Hambleton & van der Linden, 1997; Hambleton & Swaminathan, 1985). The process of determining the specific mathematical relationship between θ and p is called item calibration. After items are calibrated, they are defined by a set of parameters that specify a nonlinear, monotonically increasing relationship between θ and p . Once the item parameters are known, an estimate of θ for each student can be calculated. This estimate, $\hat{\theta}$, is considered to be an estimate of the student's true score, or a general representation of student performance. It has characteristics that may be preferable to those of raw scores for equating purposes.

For the 2013–14 CRT, the one-parameter logistic (1PL) model, which can be simplified from the three-parameter logistic (3PL) model, was used for dichotomous items (Hambleton & van der Linden, 1997; Hambleton, Swaminathan, & Rogers, 1991), and the partial credit model (PCM), which can be simplified from the generalized partial credit model, was used for polytomous items (Nering & Ostini, 2010). The 3PL model for dichotomous items can be defined as

$$P_i(1|\theta_j, \xi_i) = c_i + (1 - c_i) \frac{\exp[D a_i(\theta_j - b_i)]}{1 + \exp[D a_i(\theta_j - b_i)]}$$

where

i indexes the items,

j indexes students,

a represents item discrimination,

b represents item difficulty,

c is the pseudo guessing parameter,

ξ_i represents the set of item parameters (a , b , and c), and

D is a normalizing constant equal to 1.701.

In the case of the Montana CRT, the a_i term in the equation is equal to 1.0, and the term is equal to 0.0 for all items, which reduces to the 1PL model:

$$P_i(\theta) = \frac{\exp D(\theta - b_i)}{1 + \exp D(\theta - b_i)}$$

For polytomous items, the generalized partial credit model can be defined as

$$P_{jk}(\theta) = \frac{\exp \sum_{v=0}^k [D a_j(\theta - b_j + d_v)]}{\sum_{c=1}^m \exp \sum_{v=1}^c [D a_j(\theta - b_j + d_v)]}$$

where

j indexes items,

k indexes students,

a represents item discrimination,

b represents item difficulty,

d represents category step parameter, and
 D is a normalizing constant equal to 1.701.

In the case of the Montana CRT, the α_j term in the equation is equal to 1.0 for all items.

For more information about item calibration and determination, the reader is referred to Lord and Novick (1968), Hambleton and Swaminathan (1985), or Baker and Kim (2004).

6.2 ITEM RESPONSE THEORY RESULTS

The tables in Appendix H give the IRT item parameters of all common items on the 2013–14 CRT for science. In addition, Appendix I shows graphs of the TCCs and TIFs, which are defined below.

TCCs display the expected (average) raw score associated with each θ_j value between -4.0 and 4.0. Mathematically, the TCC is computed by summing the ICCs of all items that contribute to the raw score. Using the notation introduced in Section 6.1, the expected raw score at a given value of θ_j is

$$E(X|\theta_j) = \sum_{i=1}^n P_i(1|\theta_j)$$

where

i indexes the items (and n is the number of items contributing to the raw score),

j indexes students (here, θ_j runs from -4 to 4), and

$E(X|\theta_j)$ is the expected raw score for a student of ability θ_j .

The expected raw score monotonically increases with θ_j , consistent with the notion that students of high ability tend to earn higher raw scores than do students of low ability. Most TCCs are “S-shaped”—flatter at the ends of the distribution and steeper in the middle.

The TIF displays the amount of statistical information that the test provides at each value of θ_j . Information functions depict test precision across the entire latent trait continuum. There is an inverse relationship between the information of a test and its standard error of measurement (SEM). For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, & Rogers, 1991), as follows:

$$SEM(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}}$$

Compared to the tails, TIFs are often higher near the middle of the distribution, where most students are located.

PARSCALE v4.1 (Muraki & Bock, 2003) software was used to perform all IRT analyses for the Montana CRT. Each item occupied only one block in the calibration run, and the 1.701 normalizing constant was used. A default convergence criterion of 0.001 was used. The number of Newton cycles required for

convergence for each grade and content area during the IRT analysis can be found in Table 6-2. The number of cycles required fell within acceptable ranges.

Table 6-2. 2013–14 MontCAS: Number of Newton Cycles Required for Convergence—Science

Grade	Cycles	
	Initial	Equating
4	11	7
8	4	4
10	5	1

6.3 EQUATING

The purpose of equating is to ensure that scores obtained from different forms of a test are equivalent to each other. Equating may be used if multiple test forms are administered in the same year, as well as to equate one year’s forms to those given in the previous year. Equating ensures that students are not given an unfair advantage or disadvantage because the test form they took is easier or harder than those taken by other students.

Equating for the Montana CRT used the anchor-test-nonequivalent-groups design described by Petersen, Kolen, and Hoover (1989). In this equating design, no assumption is made about the equivalence of the examinee groups taking different test forms (that is, naturally occurring groups are assumed). IRT is particularly useful for equating nonequivalent groups (Allen & Yen, 1979). The fixed common-item IRT procedure was used. The anchor items from the previous year’s administration were identified during this year’s calibrations, and their IRT parameters were fixed to last year’s values. This method results in all person and item parameters being on the same θ scale as they were in the previous year. The procedures used for equating and scaling do not change the ranking of students, give more weight to particular items, or change students’ performance-level classifications.

6.4 EQUATING RESULTS

An Equating Report was submitted to the OPI for their approval prior to production of student reports. The Equating report details the results of a variety of quality control activities that were implemented within the Psychometrics and Research Department during IRT calibration and equating, including examining *b*-plots and TCCs and conducting delta and rescore analyses. The evaluations of the equating results are summarized in Table 6-1 above. The *b*-plots can be found in Appendix J. The procedures used to evaluate equating items are described below.

Appendix K presents the results from the delta analysis. This procedure was used to evaluate the performance of equating items, and the discard status presented in the appendix indicates whether the item was used in equating. As can be seen in the appendix, as well as in Table 6-1, a very small number of items

were identified as problematic based on the results of the delta analyses and were excluded from use in equating.

Also presented in Appendix K are the results from the rescore analysis. With this analysis, 200 random papers from the previous year were interspersed with this year’s papers to evaluate scorer consistency from one year to the next. All effect sizes were well below the criterion value for excluding an item as an equating item, 0.80 (in absolute value).

6.5 ACHIEVEMENT STANDARDS

Cutpoints for the Montana CRT in reading and mathematics were set at standard-setting meetings held in June and July 2006, and cutpoints in science were set in June 2008. Details of the standard-setting procedures can be found in the standard-setting reports and technical reports of those years. The cuts on the theta scale that were established at those meetings are presented in Table 6-3 below. The θ -metric cut scores that emerged from the standard-setting meetings will remain fixed throughout the assessment program unless standards are reset for any reason. Also shown in the table are the cutpoints on the reporting score scale (described below).

**Table 6-3. 2013–14 MontCAS: Cut Scores on the Theta Metric and Reporting Scale—
Science**

Grade	Theta			Scaled Score				
	Cut 1	Cut 2	Cut 3	Minimum	Cut 1	Cut 2	Cut 3	Maximum
4	-0.70081	-0.14474	0.55956	200	225	250	282	300
8	-0.57275	-0.07715	0.58285	200	225	250	283	300
10	-0.37793	0.12744	0.52244	200	225	250	270	300

6.5.1. Distributions

Table L-1 in Appendix L shows performance-level distributions for each of the last three years by subject and grade.

6.6 SCALED SCORES

6.6.1 Description of Scale

Montana CRT scores in each content area are reported on a scale ranging from 200 to 300. By providing information that is more specific about the position of a student’s results, scaled scores supplement performance-level scores. School- and district-level scaled scores are calculated by computing the average of student-level scaled scores. Students’ raw scores (i.e., total number of points) on the 2013–14 Montana CRT were translated to scaled scores by using a data-analysis process called *scaling*. Scaling simply converts from

one scale to another. In the same way that a given temperature can be expressed on either the Fahrenheit or Celsius scale, or the same distance can be expressed in either miles or kilometers, student scores on the 2013–14 Montana CRT tests can be expressed in raw or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change students' performance-level classifications. Given the relative simplicity of raw scores, it is fair to ask why scaled scores instead of raw scores are used in Montana CRT reports. Foremost, scaled scores offer the advantage of simplifying result reporting across content areas, grade levels, and subsequent years. Because the standard-setting process typically results in different cut scores across content areas on a raw score basis, it is useful to transform these raw cut scores to a scale that is more easily interpretable and consistent. For the Montana CRT, a score of 225 is the cut score between the Novice and Nearing Proficiency performance levels. This is true regardless of content area, grade level, or year. For example, the raw cut score between Novice and Nearing Proficiency may be 35 in grade 8 mathematics, but 33 in grade 10 mathematics. Using scaled scores greatly simplifies the task of understanding how a student performed. The raw score-to-scaled score look-up tables for science by are presented in Appendix M.

6.6.2 Calculations

For Montana CRT, scaled scores were obtained by a simple translation of students' scores using a linear equation of the form

$$SS = mY + b$$

where
 m is the slope,
 b is the intercept, and
 Y represents the student's score.

A separate linear transformation was used for each grade/content area combination. Each line was determined by using threshold values obtained via standard setting and fixing the Novice/Nearing Proficiency and Nearing Proficiency/Proficient scaled score cuts to 225 and 250, respectively. The cut between Proficient and Advanced was then allowed to vary across grades and content areas. The scaled score values obtained using this formula were rounded to the nearest integer and truncated, as necessary, so that no student received a score lower than 200 or higher than 300.

For science, the student score used for scaling was the ability estimate on the theta scale, $\hat{\theta}$, which was found from the students' raw scores by mapping through the TCC. For reading and mathematics, on the other hand, scaling was done from raw score. As with science, the students' raw scores on the 2013–14 test were transformed into ability estimates on the theta scale using the TCC. These ability estimates were then transformed into an expected raw score on the reference test form (2005–06, when standards were established for reading and mathematics) using the TCC for the reference test. This expected raw score was then scaled onto the reporting metric.

Table 6-4 shows the scaling constants for science by grade.

**Table 6-4. 2013–14 MontCAS: Scaled Score Slope and Intercept—
Science**

<i>Grade</i>	<i>Slope</i>	<i>Intercept</i>
4	44.9584	256.5073
8	50.4439	253.8917
10	49.4687	243.6957

6.6.3 Distributions

Graphs of the scaled score cumulative frequency distributions for the last three years are presented in Appendix L. Note that the graphs show the percent of students at or below each scaled score; thus, the lowest line in a given graph depicts the highest performing group. For example, in the graph for grade 4 science (Figure L-1), the line showing the cumulative distribution for 2013–14 is to the right of the line for 2012–13 which, in turn, is to the right of the line for 2011–12. This pattern indicates that student performance on the grade 5 mathematics test has improved in each of the last two years.

CHAPTER 7 RELIABILITY

Although an individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way items function together and complement one another. Tests that function well provide a dependable assessment of the student's level of ability. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student's score being either higher or lower than his or her true ability. For example, a student may misread an item or mistakenly fill in the wrong bubble when he or she knew the answer. Collectively, extraneous factors that impact a student's score are referred to as measurement error. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores, or vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors made are small on average and student scores on such a test will consistently represent their ability) are described as reliable.

There are a number of ways to estimate an assessment's reliability. One approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable (this is referred to as "test-retest reliability"). A potential problem with this approach is that students may remember items from the first administration or may have gained or lost knowledge or skills in the interim between the two administrations. A solution to the "remembering items" problem is to give a different but parallel test at the second administration. If student scores on each test correlate highly, the test is considered reliable (this is known as "alternate forms reliability," because an alternate form of the test is used in each administration). This approach, however, does not address the problem that students may have gained or lost knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms generally preclude the use of parallel forms reliability indices. One way to address the latter problem is to split the test in half and then correlate students' scores on the two half-tests; this, in effect, treats each half-test as a complete test. By doing this, the problems associated with an intervening time interval and of creating and administering two parallel forms of the test are alleviated. This is known as a "split-half estimate of reliability." If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test halves will result in a different correlation. Another problem with the split-half method of calculating

reliability is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, α (alpha), which eliminates the problem of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2013–14 Montana CRT:

$$\alpha \equiv \frac{n}{n-1} \left[1 - \frac{\sum_{i=1}^n \sigma_{(Y_i)}^2}{\sigma_x^2} \right]$$

where
i indexes the item,
n is the total number of items,
 $\sigma_{(Y_i)}^2$ represents individual item variance, and
 σ_x^2 represents the total test variance.

7.1 RELIABILITY AND STANDARD ERRORS OF MEASUREMENT

Table 7-1 presents descriptive statistics, Cronbach's α coefficient, and raw score standard errors of measurement for science by grade. (Statistics are based on common items only.)

Table 7-1. 2013–14 MontCAS: Raw Score Descriptive Statistics, Cronbach's Alpha, and Standard Errors of Measurement—Science

Grade	Number of Students	Raw Score			Alpha	SEM*
		Maximum	Mean	Standard Deviation		
4	10844	61	37.77	9.78	0.88	3.42
8	10343	61	35.94	10.26	0.87	3.67
10	10152	61	35.04	10.2	0.88	3.5

SEM = standard errors of measurement

For science, the reliability coefficients ranged from 0.87 to 0.88. Because different grades and content areas have different test designs (e.g., the number of items varies by test), it is inappropriate to make inferences about the quality of one test by comparing its reliability to that of another test from a different grade and/or content area.

7.2 SUBGROUP RELIABILITY

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2013–14 Montana CRT. Appendix N presents reliabilities for various subgroups of interest. Subgroup Cronbach's α 's were calculated using the formula defined above, based only on the members of the subgroup in question in the computations; values are only calculated for subgroups with 10 or more students. For science, subgroup reliabilities ranged from 0.68 to 0.90.

For several reasons, the results of this section should be interpreted with caution. First, inherent differences between grades and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test, but also on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix N that subgroup sample sizes may vary considerably, which results in natural variation in reliability coefficients. Additionally, α , which is a type of correlation coefficient, may be artificially depressed for subgroups with little variability (Draper & Smith, 1998). Third, there is no industry standard to interpret the strength of a reliability coefficient, and this is particularly true when the population of interest is a single subgroup.

7.3 REPORTING SUBCATEGORY RELIABILITY

Of even more interest are reliabilities for the reporting subcategories within Montana CRT content areas, described in Chapter 3. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Appendix N. Once again, as expected, because they are based on a subset of items rather than the full test, computed subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, and interpretations should take this into account.

For science, subcategory reliabilities ranged from 0.24 to 0.69. The subcategory reliabilities were lower than those based on the total test and approximately to the degree one would expect based on classical test theory. Qualitative differences between grades and content areas once again preclude valid inferences about the quality of the full test based on statistical comparisons among subtests.

7.4 INTERRATER CONSISTENCY

Chapter 4 of this report describes in detail the processes that were implemented to monitor the quality of the hand-scoring of student responses for short-answer and constructed-response items. One of these processes was double-blind scoring: approximately 2% of student responses were randomly selected and scored independently by two different scorers. Results of the double-blind scoring were used during scoring to identify scorers who required retraining or other intervention and are presented here as evidence of the reliability of the Montana CRT. A summary of the interrater consistency results is presented in Table 7-2 below. Results in the table are collapsed across the hand-scored items by grade and number of score categories (two for short-answer items and five for constructed-response items). The table shows the number of included scores, the percent exact agreement, the percent adjacent agreement, the correlation between the first two sets of scores, and the percent of responses that required a third score. This same information is provided at the item level in Appendix O.

Table 7-2. 2013–14 MontCAS: Summary of Interrater Consistency Statistics Collapsed across Items—Science

Grade	Number of		Percent		Correlation	Percent of Third Scores
	Score Categories	Included Scores	Exact	Adjacent		
4	5	439	66.51	28.93	0.81	4.56
8	5	441	59.86	34.47	0.84	5.67
10	5	389	73.78	22.88	0.87	3.34

7.5 RELIABILITY OF PERFORMANCE-LEVEL CATEGORIZATION

While related to reliability, the accuracy and consistency of classifying students into performance categories are even more important statistics in a standards-based reporting framework (Livingston & Lewis, 1995). After the performance levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For the Montana CRT, students are classified into one of four performance levels: Novice (N), Nearing Proficiency (NP), Proficient (P), or Advanced (A). This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated because errorless test scores do not exist. Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2013–14 Montana CRT because it is easily adaptable to all types of testing formats, including mixed format tests.

The accuracy and consistency estimates reported in Appendix P make use of “true scores” in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. Of course, true scores cannot be observed and so must be estimated. In the Livingston and Lewis method, estimated true scores are used to categorize students into their “true” classifications.

For the 2013–14 Montana CRT, after various technical adjustments (described in Livingston & Lewis, 1995), a four-by-four contingency table of accuracy was created for science by grade, where cell $[i, j]$ represented the estimated proportion of students whose true score fell into classification i (where $i = 1$ to 4) and whose observed score fell into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students whose true and observed classifications matched) signified overall accuracy.

To calculate consistency, true scores were used to estimate the joint distribution of classifications on two independent, parallel test forms. Following statistical adjustments, per Livingston and Lewis (1995), a new four-by-four contingency table was created for science by grade and populated by the proportion of students who would be categorized into each combination of classifications according to the two (hypothetical) parallel test forms. Cell $[i, j]$ of this table represented the estimated proportion of students whose observed score on the first form would fall into classification i (where $i = 1$ to 4) and whose observed score on the second form would fall into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students categorized by the two forms into exactly the same classification) signified overall consistency.

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{(\text{Observed agreement}) - (\text{Chance agreement})}{1 - (\text{Chance agreement})} = \frac{\sum_i C_{ii} - \sum_i C_{i.} C_{.i}}{1 - \sum_i C_{i.} C_{.i}}$$

where

$C_{i.}$ is the proportion of students whose observed performance level would be Level i (where $i = 1-4$) on the first hypothetical parallel form of the test;

$C_{.i}$ is the proportion of students whose observed performance level would be Level i (where $i = 1-4$) on the second hypothetical parallel form of the test;

C_{ii} is the proportion of students whose observed performance level would be Level i (where $i = 1-4$) on both hypothetical parallel forms of the test.

Because κ is corrected for chance, its values are lower than other consistency estimates.

7.5.1 Decision Accuracy and Consistency Results

The decision accuracy and consistency analyses described above are provided in Table P-1 of Appendix P. The table includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon performance level are also given. For these calculations, the denominator is the proportion of students associated with a given performance level. For example, the conditional accuracy value is 0.77 for Novice for science grade 4. This figure indicates that among the students whose true scores placed them in this classification, 77 percent would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.78 indicates that 78 percent of students with observed scores in the Novice level would be expected to score in this classification again if a second, parallel test form were used.

For some testing situations, the greatest concern may be decisions around level thresholds. For example, in testing done for No Child Left Behind accountability purposes, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. For the 2013–14 Montana CRT, Table P-2 in Appendix P provides accuracy and consistency estimates at each cutpoint, as well

as false positive and false negative decision rates. A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.

The above indices are derived from Livingston and Lewis's (1995) method of estimating the accuracy and consistency of classifications. It should be noted that Livingston and Lewis discuss two versions of the accuracy and consistency tables. A standard version performs calculations for forms parallel to the form taken. An "adjusted" version adjusts the results of one form to match the observed score distribution obtained in the data. The tables use the standard version for two reasons: (1) this "unadjusted" version can be considered a smoothing of the data, thereby decreasing the variability of the results; and (2) for results dealing with the consistency of two parallel forms, the unadjusted tables are symmetrical, indicating that the two parallel forms have the same statistical properties. This second reason is consistent with the notion of forms that are parallel; that is, it is more intuitive and interpretable for two parallel forms to have the same statistical distribution.

Descriptive statistics relating to the decision accuracy and consistency (DAC) of the 2013–14 Montana CRT tests can be derived from Table P-1. For science, overall accuracy ranged from 0.91 to 0.96, overall consistency ranged from 0.88 to 0.95, and the kappa statistic ranged from 0.52 to 0.55. Note that, as with other methods of evaluating reliability, DAC statistics calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Appendix P should be interpreted with caution. In addition, it is important to remember that it is inappropriate to compare DAC statistics between grades and content areas.

CHAPTER 8 SCORE REPORTING

The Montana CRT is designed to measure student performance against Montana’s content standards. Consistent with this purpose, results on the CRT were reported in terms of performance levels that describe student performance in relation to these established state standards. There are four performance levels: Novice, Nearing Proficiency, Proficient, and Advanced (performance-level distributions are given in Appendix L). Students receive a separate performance-level classification, based on total scaled score, in each content area.

State results were provided to the OPI via a secure Web site. Science reporting data for the 2013–14 Montana CRT were made available to systems and schools online via the Montana Analysis and Reporting System (MARS) on June 3, 2014. Student reports were delivered to System Test Coordinators for distribution to parents on September 5, 2014. Student reports were also posted online to be accessible to schools. System test coordinators and teachers were also provided with copies of the *Guide to the 2014 Criterion-Referenced Test and CRT-Alternate Assessment Reports* to assist them in understanding the connection between the assessment and the classroom. The guide provides information about the assessment and the use of assessment results.

School- and system-level results are reported as the number and percentages of students attaining each performance level at each grade level tested. As described below, decision rules were formulated in early 2014 by the OPI and Measured Progress to identify students who, during the reporting process, were to be excluded from school- and system-level reports. (A copy of these decision rules is included in this report as Appendix A.) State-level summary reports were also produced.

The reports described in the sections that follow are separated into two categories. The first set of reports described is static reports, which are provided online as PDF documents; student reports are also provided on paper. The static reports are the following:

- student report (paper and online)
- school, system, and state summary reports (online)

The remaining reports are interactive reports, provided via MARS (see Sections 9.3 and 9.4 below):

- class roster and item-level reports
- performance-level summary
- released items summary data
- longitudinal data report

Sample Report Shells are included as Appendix Q.

8.1 DECISION RULES

As mentioned above, to ensure that reported results for the 2013–14 Montana CRT are accurate relative to collected data and other pertinent information, a document that delineates analysis and reporting rules was created. These decision rules were observed in the analyses of Montana CRT test data and in reporting the test results. Moreover, these rules are the main reference for quality-assurance checks.

The decision rules document used for reporting results of the 2014 administration of the Montana CRT is found in Appendix A.

The rules primarily describe the inclusion/exclusion of students at the school, system, and state levels of aggregation. The document also describes rules as they pertain to individual reports. Finally, it describes the classification of students based on their school type or other information provided by the state through the student demographic file (AIM) or collected on the answer booklet.

8.2 STATIC REPORTS

8.2.1 Student Report

The student report is produced for each parent of a student who took or was eligible to take the Montana CRT. The report is shipped to systems and posted online for school/system access.

The student report gives the results for grades 4, 8, and 10 for science and gives the earned performance level and scaled score for each subject. The report also provides a comparison of the student's performance to that of the state as a whole for science. The report contains the results for science at the content-standard level. The number of points earned by the student in each content standard is reported, as is the range of points earned by students who achieve proficiency.

8.2.2 Summary Reports

The summary report is produced at the school, system, and state levels. The report is produced for science grades 4, 8, and 10. The report consists of three sections: Distribution of Scores, Subtest Results, and Results for Subgroups of Students.

The Distribution of Scores section of the report contains a breakdown of the performance of included students (as described in the decision rules document) into different scaled score intervals. The number and percent of students that fall into each scaled score interval is shown. There is an overall percentage reported for students that fall into each of the four performance levels (Novice, Nearing Proficiency, Proficient, and Advanced). In the School Summary Report, the calculations are done at the school, system, and state levels. The System Summary Report contains results at the system and state levels. The State Summary Report contains only state-level results.

The Subtest Results section of the report summarizes the average points earned in the different content standards by included students (as described in the decision rules document) in the school, system, and state. The average points earned are compared to the total possible points for each content standard.

The Results for Subgroups of Students section of the report summarizes the performance of included students (as described in the decision rules document) broken down by various reporting categories. For each reporting category, the number of tested (included) students is reported, as is the percentage of students in each of the four performance levels. In the School Summary Report, this is reported at the school, system, and state levels. In the System Summary Report, the data are reported at the system and state levels. In the State Summary Report, the data are reported at the state level only.

The reporting categories are as follows:

- All Students
- Gender
- Ethnicity (American Indian or Alaska native, Asian, Hispanic, Black or African American, Native Hawaiian or Other Pacific Islander, White)
- Special Education
- Students with a 504 Plan
- Title I (optional)
- Tested with Standard Accommodation
- Tested with Nonstandard Accommodation
- Alternate Assessment (results are not given for this category on the Montana CRT Summary reports)
- Migrant
- Gifted/Talented
- LEP/ELL
- Former LEP Student
- LEP Student Enrolled for First Time in a U.S. School
- Free/Reduced Lunch

Data are suppressed if there are less than 10 students tested (included) in a reporting category at a given aggregation level.

The data for the reporting categories were provided by information coded on the students' answer booklets by teachers and/or data supplied by the state through an AIM export. Due to relatively low numbers of students in certain reporting categories, school personnel are advised, under FERPA guidelines, to treat these pages confidentially.

8.3 MONTANA ANALYSIS AND REPORTING SYSTEM

Using advanced Web technology, MARS gives Montana educators and administrators the ability to filter data based on test year, grade level, content area, standard, and student subgroup. This allows administrators to isolate cross-sections of the results and identify areas of strong or poor performance.

The confidential nature of the data in MARS necessitates the strict enforcement of site security. All transmissions are done over Secure Socket Layers (SSL). A system of user role definitions and permissions dictates the scope of access granted to individual users. Organizations (system or school levels) are given administrative power to grant or deny access to their data within the system, and they have the ability to disable users. Personnel using MARS may be granted permission to view students' results at an organizational level or only a select group as defined by the administrator. Predefined reports are included in the system, as is the ability to render and print additional copies.

8.3.1 User Accounts

In MARS, principals have the ability to create unique user accounts by assigning specific usernames and passwords to educators in their school, such as teachers, curriculum coordinators, or special education coordinators. Once the accounts have been created, individual students may be assigned to each user account. After users have received their usernames and passwords, they are able to log in to their accounts and access the interactive reports, which will be populated only with the subgroup of students assigned to them.

Information about the interactive reports and setting up user accounts is available in the *Analysis & Reporting System User Manual* that is available for download on the MARS system.

8.4 INTERACTIVE REPORTS

As mentioned above, there are four interactive reports that are available from MARS: Roster Report, Performance-Level Summary, Released Items Summary Data, and Longitudinal Data. Each of these interactive reports is described in the following sections. Sample interactive reports are provided in Appendix Q. To access these four interactive reports, the user clicks the interactive tab on the home page of the system and selects the report desired from the drop down menu. Next, the user applies basic filtering options, such as the name of the district or school and the grade level/content area test, to open the specific report. At this point, the user has the option of printing the report for the entire grade level or applying advanced filtering options to select a subgroup of students to analyze. Advanced filtering options include gender, ethnicity, limited English proficient, Individual Education Program, migrant, and 504 plan. All interactive reports, with the exception of the Longitudinal Data Report, allow the user to provide a custom title for the report.

8.4.1 Roster Report

The Montana CRT Roster Report provides a roster of all students in a school and provides performance on the common items that are released to the public, one report per content area. For all grades, the student names and identification numbers are listed as row headers down the left side of the report. The items are listed as column headers in the same order they appeared in the released item document.

For each item, the following are shown:

- depth of knowledge code
- item type
- correct response key for multiple-choice items
- total possible points
- content standard

For each student, multiple-choice items are marked either with a plus sign (+), indicating that the student chose the correct multiple-choice response, or a letter (from A to D), indicating the incorrect response chosen by the student. For short-answer and constructed-response items, the number of points earned is shown. All responses to released items are shown in the report, regardless of the student's participation status. The columns on the right side of the report show the total test results, broken into several categories. Subcategory points earned columns show points earned by the student in each content area subcategory relative to total possible points. The total points earned column is a summary of all points earned and total possible points in the content area. The last two columns show the student's scaled score and performance level. Students reported as Not Tested are given a code in the performance-level column to indicate the reason the student did not test. It is important to note that not all items used to compute student scores are included in this report, only released items. At the bottom of the report, the average percentage correct for each multiple-choice item and average scores for the short-answer and constructed-response items are shown for the school, district, and state. When advanced filtering criteria are applied by the user, the school and district percent correct/average score rows at the bottom of the report are blanked out and only the group row and the state row for the group selected will contain data. This report can be saved, printed, or exported as a PDF.

The Montana CRT roster is confidential and should be kept secure within the school and district. FERPA requires that access to individual student results be restricted to the student, the student's parents/guardians, and authorized school personnel.

8.4.2 Performance-Level Summary

The Performance-Level Summary provides a visual display of the percentages of students in each performance level for a selected grade/content area. The four performance levels (Novice, Nearing Proficiency, Proficient, and Advanced) are represented by various colors in a pie chart. A separate table is

also included below the chart that shows the number and percentage of students in each performance level. This report can be saved, printed, or exported as a PDF or JPG file.

8.4.3 Item Analysis Data

The Released Items Summary Data report is a school-level report that provides a summary of student responses to the released items for a selected grade/content area. The report is divided into two sections by item type (multiple-choice and open-response). For multiple-choice items, the total number/percent of students who answered the item correctly and the number of students who chose each incorrect option or provided an invalid response are reported. An invalid response on a multiple-choice item is defined as “the item was left blank” or “the student selected more than one option for the item.” For open-response items, point value and average score for the item are reported. Users are also able to view the actual released items within this report. If a user clicks on a particular magnifying glass icon next to a released item number, a pop-up box will open displaying the released item.

8.4.4 Longitudinal Data Report

The longitudinal data report is a confidential student-level report that provides individual student performance data for multiple test administrations. Results are reported for a student going back to academic year 2006–07. The state-assigned student identification number is used to link students across test administrations. Student performance on future test administrations will be included on this report over time. This report can be saved, printed, or exported as a PDF file.

8.5 INTERPRETIVE MATERIALS AND WORKSHOPS

An interpretive guide to the CRT reports is provided on the OPI Web site: <http://opi.mt.gov/>.

8.6 QUALITY ASSURANCE

Quality-assurance measures are embedded throughout the entire process of analysis and reporting. The data processor, data analyst, and psychometrician assigned to work on the Montana CRT implement quality-control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different functions within the Data Services and Static Reporting and Psychometrics and Research departments, the sending function verifies that the data are accurate before handoff. Additionally, when a function receives a data set, the first step is to verify the data for accuracy.

Another type of quality-assurance measure is parallel processing. Different exclusions that determine whether each student receives scaled scores and/or is included in different levels of aggregation are parallel processed. Using the decision rules document, two data analysts independently write a computer program that

assigns students' exclusions. For each content area and grade combination, the exclusions assigned by each data analyst are compared across all students. Only when 100% agreement is achieved can the rest of data analysis be completed.

Another level of quality assurance involves the procedures implemented by the quality-assurance group to check the accuracy of reported data. Using a sample of schools and districts, the quality-assurance group verifies that reported information is correct. The step is conducted in two parts: (1) verify that the computed information was obtained correctly through appropriate application of different decision rules, and (2) verify that the correct data points populate each cell in the Montana CRT reports. The selection of sample schools and districts for this purpose is very specific and can affect the success of the quality control efforts. There are two sets of samples selected that may not be mutually exclusive.

The first set includes those that satisfy the following criteria:

- one-school district
- two-school district
- multi-school district

The second set of samples includes districts or schools that have unique reporting situations, as indicated by decision rules. This second set is necessary to ensure that each rule is applied correctly. The second set includes those that satisfy the following criteria:

- private school
- school with excluded (not tested) students

The quality assurance group uses a checklist to implement its procedures. After the checklist is completed, sample reports are circulated for psychometric checks and program management review.

CHAPTER 9 VALIDITY

Because interpretations of test scores—and not a test itself—are evaluated for validity, the purpose of the *2013–14 Montana CRT Technical Report* is to describe several technical aspects of the Montana CRT tests in support of score interpretations (AERA, 1999). Each chapter contributes an important component in the investigation of score validation: test development and design; test administration; scoring, scaling, and equating; item analyses; reliability; and score reporting.

As stated in the overview chapter, *Standards for Educational and Psychological Testing* (AERA et al., 1999) provides a framework for describing sources of evidence that should be considered when constructing a validity argument. The evidence around test content, response processes, internal structure, relationship to other variables, and consequences of testing speak to different aspects of validity but are not distinct types of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

Evidence on test content validity is meant to determine how well the assessment tasks represent the curriculum and standards for each content area and grade level. Content validation is informed by the item-development process, including how the test blueprints and test items align to the curriculum and standards. Viewed through this lens provided by the Standards, evidence based on test content was extensively described in Chapters 2 and 3. Item alignment with Montana content standards; item bias, sensitivity, and content-appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test-administration training are all components of validity evidence based on test content. As discussed earlier, all CRT questions are aligned by Montana educators to specific Montana content standards and undergo several rounds of review for content fidelity and appropriateness. Items are presented to students in multiple formats (constructed-response and multiple-choice). Finally, tests are administered according to state-mandated standardized procedures, with allowable accommodations, and all test proctors are required to attend annual training sessions.

The scoring information in Chapter 4 describes the steps taken to train and monitor hand-scorers, as well as quality-control procedures related to scanning and machine scoring. To speak to student response processes, however, additional studies would be helpful and might include an investigation of students' cognitive methods using think-aloud protocols.

Evidence based on internal structure is presented in great detail in the discussions of item analyses, reliability, and scaling and equating in Chapters 5 through 7. Technical characteristics of the internal structure of the assessments are presented in terms of classical item statistics (item difficulty, item-test correlation), differential item functioning analyses, dimensionality analyses, reliability, standard errors of measurement, and item response theory parameters and procedures. Each test is equated to the same grade and content-area

test from the prior year in order to preserve the meaning of scores over time. In general, item difficulty and discrimination indices were in acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall.

Evidence based on the consequences of testing is addressed in the scaled-scores information in Chapter 6 and the reporting information in Chapter 8, as well as in the test-interpretation guide, which is a separate document that is referenced in the discussion of reporting. Each of these chapters speaks to the efforts undertaken to promote accurate and clear information provided to the public regarding test scores. Scaled scores offer the advantage of simplifying the reporting of results across content areas, grade levels, and subsequent years. Performance levels provide users with reference points for mastery at each grade level, which is another useful and simple way to interpret scores. Several different standard reports are provided to stakeholders. In addition, a data-analysis tool is provided to each school system to allow educators the flexibility to customize reports for local needs. Additional evidence of the consequences of testing could be supplemented with broader investigation of the impact of testing on student learning.

To further support the validation of the assessment program, additional studies might be considered to provide evidence regarding the relationship of CRT results to other variables, including the extent to which scores from the CRT converge with other measures of similar constructs and the extent to which they diverge from measures of different constructs. Relationships among measures of the same or similar constructs can sharpen the meaning of scores and appropriate interpretations by refining the definition of the construct.

The evidence presented in this report supports inferences of student achievement on the content represented on the Montana content standards for reading, mathematics, and science for the purposes of program and instructional improvement and as a component of school accountability.

REFERENCES

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth, Inc.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baker, F. B., & Kim, S-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker, Inc.
- Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). Fort Worth: Holt, Rinehart and Winston.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley and Sons, Inc.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.
- Hambleton, R. K., & van der Linden, W. J. (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.
- Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Joint Committee on Testing Practices. Available from www.apa.org/science/programs/testing/fair-code.aspx
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Muraki, E., & Bock, R. D. (2003). *PARSCALE 4.1*. Lincolnwood, IL: Scientific Software International.

- Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Routledge.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262).
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *52*, 589–617.
- Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M., A. J. van Duign, & T. A. B. Snijders (Eds.), *Essays on item response theory*, (pp. 357–375). New York: Springer-Verlag.
- Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

APPENDICES

APPENDIX A—ANALYSIS AND REPORTING DECISION RULES

**Analysis and Reporting Decision Rules
Montana Comprehensive Assessment System (MontCAS) CRT and CRT-Alternate
Spring 13-14 Administration**

This document details rules for analysis and reporting. The final student level data set used for analysis and reporting is described in the “Data Processing Specifications.” This document is considered a draft until the Montana Office of Public Instruction (OPI) signs off. If there are rules that need to be added or modified after said sign-off, OPI sign off will be obtained for each rule. Details of these additions and modifications will be in the Addendum section.

I. General Information
A. Tests Administered

Grade	Subject	Items included in Raw Score		IABS Reporting Categories (Standards) (Not Applicable for CRT-Alternate)
		CRT	CRT-Alt	
03	Reading Math	Not Tested	All	Cat2
04	Reading Math	Not Tested	All	Cat2
	Science	Common	All	Cat3
05	Reading Math	Not Tested	All	Cat2
06	Reading Math	Not Tested	All	Cat2
07	Reading Math	Not Tested	All	Cat2
08	Reading Math	Not Tested	All	Cat2
	Science	Common	All	Cat3
10	Reading Math	Not Tested	All	Cat2
	Science	Common	All	Cat3

- B. Reports Produced
1. Student Labels (Printed)
 2. Student Report (Printed and posted online)
 3. Roster & Item Level Report (CRT-Alt: posted online; CRT:Interactive System)
 - by grade, subject and class/group
 4. Summary Report (Online)
 - Consists of sections:

- I. Distribution of Scores
 - II. Subtest Results
 - III. Results for Subgroups of Students
 - by grade, subject and school
- by grade, subject and system
- State summary reports are not produced
- The summary reports will be named as described below. This naming convention allows unique names for each PDF generated.
- [Contract Nick Name][Report Name][Grade][Subject]_[District/School Code].pdf
- Where
- Contract Nick Name - Montana1314, MTAlt1314
- Report Name - SummarySystem, SummarySchool
- Grade - 03-08, 10
- Subject – Mat*, Rea*, Sci (*Alt only)

C. Files Produced

1. One state file for each grade (Format: comma delimited format)
 - a. Consists of student level results
 - b. Alternately assessed students are in separate files by grade.
 - c. Naming conventions
 - i. CRT All subjects- Studentdatafile[2 digit grade].csv
 - ii. CRT-Alternate All subjects- altStudentdatafile[2 digit grade].csv
 - d. File layout: Studentdatafilelayout.xls and altstudentdatafilelayout.xls
2. System level files (Format: Excel ; Online)
 - a. Consists of student level results for each system for each grade. Contains all subjects tested at that grade.
 - b. Naming convention: Studentdatafile[2 digit grade].xls
 - c. File Layout: Systemstudentdatafilelayout.xls
3. School level file (Format: Excel; Online)
 - a. Consists of student level results for each school and grade. Contains all subjects tested at that grade.
 - b. Naming convention: Studentdatafile[2 digit grade].xls
 - c. File Layout: Systemstudentdatafilelayout.xls
4. State Student Datafiles files (Format: comma delimited format)
 - a. Consists of student level results and test metadata for the current year.
 - b. Contains all students included in CRT state files.
 - c. Naming conventions
 - i. Rawdata.csv

- ii. Scoreddata.csv
- iii. Plusdata.csv
- iv. Testmetadata.csv
- d. File layout: Rawdatalayout.xls, Scoreddatalayout.xls, Plusdatalayout.xls, Testmetadatalayout.xls

D. School Type

Schtype	Source	Description	Included in Aggregations		
			School	System	State
“Pras”	Data file provided by state	Private Accredited School. They are their own system	Yes. Same information for school & system but both sets of reports produced	Yes. Same information for school & system but both sets of reports produced	No
“Prnas”	Data file provided by state	Private non-accredited school. They are their own system	Yes. Same information for school & system but both sets of reports produced	Yes. Same information for school & system but both sets of reports produced	No
“SNE”	Scanned data/ updated by OPI	Student not enrolled	No.	No.	No.
“Oth”		Non-private school	Yes	Yes	Yes

E. Other Information

1. CRT are constructed with a combination of common and embedded field test items.
2. The CRT-Alternate consists of a set of 5 performance tasklets. The number of items in each tasklet varies.
3. Braille Students:
 - a. See Appendix A.1 for a list of the items not included in the Braille form.
 - b. If a student is identified as taking the Braille test, these items are not included in the student’s raw score. The student is scaled on a separate form based on the items that are available to him or her. See the Calculations section for more information.

II. Student Participation/Exclusions

A. Test Attempt Rules

1. A valid response to a multiple choice item is A, B, C, or D. An asterisk (multiple marks) is not considered a valid response. A valid score for an open response item is a non-blank score.
2. Incomplete (CRT): The student has exactly one (1) valid response to common items.
3. Incomplete (CRT-Alternate): The student has fewer than three (3) scores across all tasklets.
4. The student is classified as Did Not Participate (DNP) in CRT if the student does not have any valid responses for that subject in either CRT or CRT-Alternate and has no not tested reason.

B. Not Tested Reasons

1. If a student is marked First year LEP regardless of items attempted the student is considered first year LEP for reporting purposes. Reading is optional for first year in U.S schools LEP students.

C. Student Participation Status

1. The following students are excluded from all aggregations.
 - a. Foreign Exchange Students (FXS).
 - b. Homeschooled students (schtype='SNE').
 - c. Student in school less than 180 hours (PSNE).
 - d. DNP (for that subject)
 - e. First year in U.S schools LEP*(regardless of how many items were attempted)
 - f. CRT only: Student tested with Non-Standard Accommodations (NSA for that subject)*
 - g. Alt (Alt='1')

* These students are aggregated on the Disaggregated report in their respective rows.
2. If any of the non-standard accommodations are bubbled the student is considered tested with non-standard accommodations (NSA) in that subject.
3. If the student has not been in that school for the entire academic year the student is excluded from school level aggregations (NSAY).
4. If the student has not been in that system for the entire academic year the student is excluded from system and school level aggregations (NDAY).
5. If the student took the alternate assessment the student is not counted as participating in the general assessment. Alternate Assessment students receive their results on an Alternate Assessment Student Report. They are reported according to participation rules stated in this document.
6. (CRT-Alternate) If the teacher halted the administration of the assessment after the student scored zero (0) for three (3) consecutive items within tasklets , the student is classified as

Halted in that subject. If the student was halted within a tasklet then the rest of the items within the tasklet are blanked out and do not count toward the student's score. If the other tasklets are complete then those items will be counted toward the student's score.

7. If the student took the Braille form of the test the raw scores are not included in raw score school, system or state averages. They are not included in group averages on the interactive roster.

D. Student Participation Summary

Participation Status	Part. Flag	Raw score	Scaled Score	Perf. level	Included on Roster	Included in aggregations		
						Sch	Sys	Sta
FXS	E	✓	✓	✓				
SNE	E	✓	✓	✓				
PSNE	E	✓	✓	✓				
NSA(by subject) Applies to CRT only	A	✓	✓	✓	✓	Only included in count and percents on Disaggregated report for nonstandard accommodations.		
First year in U.S schools LEP	A	✓	See Report Specific Rules	See Report Specific Rules	✓			
NSAY only	B	✓	✓	✓	✓		✓	✓
NDAY	C	✓	✓	✓	✓			✓
ALT*	A	✓	✓	✓	✓	See footnote below		
Incomplete	A	✓	✓	✓	✓			
DNP (Non-Participants)	F	✓	✓	✓	✓			
Halted(CRT-Alt only by subject)	D	✓	✓	✓	✓	✓	✓	✓
Tested	Z	✓	✓	✓	✓	✓	✓	✓

* They are included in summary data only for alternate assessment reports (according to participation rules).

If a student has conflicting participation statuses the following hierarchy is applied to determine how the student is reported:

- F (Student attempted no items and is not alt and cannot be classified as first-year LEP)
- E (FXS, SNE or PSNE)
- A (NSA, first year in U.S schools LEP, ALT or INC)
- C (NDAY)
- B (NSAY)
- D (Halted; applies to CRT-Alt only)
- Z (completed CRT or CRT-Alt and none of the above conditions apply)

III. Calculations

A. Raw Scores

1. (CRT) Raw scores are calculated using the scores on common multiple choice and open response items.
2. (CRT-Alternate) Raw score is the sum of the individual item scores.

B. Scaling

1. Scaling is accomplished by defining the unique set of test forms for each grade/subject combination. This is accomplished as follows:
 - a. Translate each form and position into the unique item number assigned to the form/position.
 - b. Order the items by
 - I. Type- multiple choice, short-answer, constructed-response
 - II. Form-common, then by ascending form number.
 - III. Position
 - c. If an item number is on a form, then set the value for that item number to '1', otherwise set to '.'. Set the exception field to '0' to indicate this is an original test form.
 - d. If an item number contains an 'X' (item is not included in scaling) then set the item number to '.'. Set the exception field to '1' to indicate this is not an original test form.
 - e. Compress all of the item numbers together into one field in the order defined in step II to create the test for the student.
 - f. Select the distinct set of tests from the student data and order by the exception field and the descending test field.
 - g. Check to see if the test has already been assigned a scale form by looking in the daScaleForm table. If the test exists then assign the existing scale form. Otherwise assign the next available scale form number. All scale form numbering starts at 01 and increments by 1 up to 99.
2. Psychometrics provides a lookup table for each scale form. These lookup tables are used to assign scaled scores, performance levels and standard errors.

3. The scaled score cuts for all three subjects and all grades have been fixed and are the same as last year for the CRT.
4. Students excluded from aggregations at the state level are excluded from psychometric files.

C. CRT-Alternate: The classcode is created using the following steps:

1. The following students are not included when creating the class codes.
 - SNE
 - FXS
 - PSNE
2. The dataset (by grade) is sorted by schcode and class/group name
3. The records are then numbered consecutively starting at 1. This number is then padded with zeros (in front) to create a 3 digit number.

D. Performance Level coding:

Numeric Performance Level	Performance level Name	Abbreviation
1(lowest)	Novice	N
2	Nearing Proficiency	NP
3	Proficient	P
4(highest)	Advanced	A

E. Rounding Table

Calculation	Rounded (to the nearest)
Static Reports: Percents and averages	Whole number
Item averages : Multiple choice items	The average is multiplied by 100 and rounded to the nearest whole number.
Item averages: Open response items	Open-response item averages are rounded to the nearest tenth.

F. Minimum N size

1. The number of included students (N) in a subject is the number of students in the school/system/state minus FXS minus PRAS minus PRNAS minus PSNE minus SNE minus First year LEP minus Incomplete minus NSA minus DNP.
2. Minimum N size is 10.

3. School/system reports are produced regardless of N-size, except no reports are generated if N=0.
- G. The common items are used in reporting the average number of points for each standard.
- H. Assignment of rperflvel
1. If the student is marked as taking the CRT-Alt then rperflvel='A', otherwise
 2. If the student is classified as did not participate (DNP) then rperflvel='D', otherwise
 3. If the student is Incomplete in a subject and not marked first year LEP rperflvel='I', otherwise
 4. If the student is incomplete in Reading or has not attempted any items in Reading and is marked first year LEP rperflvel='L' for all subjects, otherwise
 5. If the student does not meet any of the above conditions then rperflvel=perflvel.

IV. Report Specific Rules

A. Student Label

1. If a student is First year LEP and incomplete in Reading, the Reading performance level is 'LEP'. The reading scaled score is blank.
2. If a student is First year LEP, the math and science performance levels are the name of the earned performance level and the scaled scores are the student's earned score.
3. If the student is not first year LEP, the performance level name corresponding to the student's earned score is displayed.
4. If the student is First year LEP but is not incomplete in Reading then the student receives his earned scaled score and performance level.
5. If the student is DNP the student receives a student label. The student receives scaled score =200 and performance level=Novice.
6. The student's name is formatted as Lname, Fname.
7. The student's name is uppercase.
8. The school and system names are title case.
9. The labels are sorted alphabetically by Lname, Fname within school and grade.
10. Test date is 2014.
11. Performance level name from section III.D above is shown on the label if the student receives a performance level.

B. Student Report

1. State performance will always appear on the student report, regardless of the student's status.
 - a. A bar on the student report will indicate the percentage of students who appear in each performance level for each subject.

2. If a student is First year LEP and incomplete in Reading, the student will receive the note "Student is Limited English Proficient (LEP). Your student is in his or her first year in a United States school. For further information please contact your school principal or testing director."
3. If the student is First year LEP but is not incomplete in Reading then the student receives his earned scaled score and performance level.
4. If a student is First year LEP, the math and science performance levels are the name of the earned performance level and the scaled score is the student's earned score.
5. If the student is not first year LEP, the performance level name corresponding to the student's earned score is displayed.
6. If the student is incomplete the student receives the scores with the note "Your student did not complete the 2014 CRT. For further information please contact your school principal or testing director."
7. If the student is NSA the student receives his scores with the note "Your student was administered the 2014 CRT with a non-standard testing accommodation. For further information please contact your school principal or testing director."
8. If there is no last name or first name for the student, the name displayed is "Name Not Provided".
9. Alt students who are halted receive their scores and performance level and the note "Teacher halted the administration of one or more of the five tasklets after the student scored a 0 for three consecutive items within a tasklet on two different test administrations. Any completed tasklets have been scored and are reflected in the student's scaled score."
10. If the student is DNP the student receives a Student Report. The student receives scaled score =200 and performance level =Novice. The standards will not be reported. The student receives the note "Student did not participate."
11. If the student had a testing irregularity the student receives the note "A test administration irregularity has affected your student's results. For further information please contact your school principal or testing director."
12. Total Points Possible, Student percent of points earned, and Average state percent are suppressed for students who took Braille test (Braille='1') or who used JAWS (JAWS='1'). This suppression is applied only to the standards which contain the items not on the student's form.
13. For each scored subject, the student report will show a bar with the subject scaled score, as well as an error bar showing the low and high scaled scores, adjusted so these scores are equidistant from the scaled score.
14. Only content standards that apply to the student are printed.
15. The following standards are not reported for either CRT or CRT-Alt:
 - a. Reading standard 3
 - b. Mathematics standard 1
 - c. Science standards 5 and 6
16. (Alt only) Do not suppress standard data regardless of the number of total possible points.

17. (Alt only) Given aggregate data are at the state level only, data are not suppressed based on total number of students.

C. Roster & Item Level Report-Alternate Assessment only

1. If a student is First year LEP and the student is not incomplete in Reading:
 - a. The math (and science) performance level is the abbreviation of the earned performance level and the scaled score is the student's earned score.
 - b. The reading performance level is the abbreviation of the earned performance level and the scaled score is the student's earned score.
 - c. The student is excluded from Reading, Math and Science aggregations.
2. If the student is First year LEP and incomplete in Reading
 - a. The student's Reading, Math (and Science) performance levels are 'LEP'
 - b. The student's math (and science) scaled score is the student's earned scaled score and the reading scaled score is blank.
 - c. The student's responses for all subjects are displayed.
 - d. The student is excluded from Math, Reading (and Science) aggregations.
3. If the student is not first year LEP, the performance level abbreviation corresponding to the student's earned score is displayed.
4. If the student is incomplete the student receives the scores with a footnote (†) "Student did not complete the assessment."
5. There is no last name or first name for the student, the name displayed is "Name Not Provided". These students appear at the bottom of the roster.
6. If class/group information is missing the roster is done at the school level.
7. Results for Alternate Assessment students are reported only on their class/group/school's alternate *Roster & Item Level Report*.
8. Within each demonstration school the class is 'DEM'.
9. Only the standards reported on the Summary report are reported on the roster.
10. The student's are sorted by lname, fname
11. Student names are formatted Lname, Fname.
12. Student names are uppercase.
13. Performance level abbreviation from section III.D is placed the performance level column if the student receives a performance level.
14. If the student is NSAY='1' or NDAY='1' then the appropriate footnote is placed beside the first name. ¥ "Not in school and/or system for full academic year."
15. If [subject]halted='1' for any subject then the appropriate footnote is placed beside the first name. § "Teacher halted the administration of one or more of the five tasklets after the student scored a 0 for three consecutive items within a tasklet on two different test administrations.

Any completed tasklets have been scored and are reflected in the student's scaled score."

16. Data are not suppressed regardless of the number of students included.
17. Standard data are not suppressed regardless of the number of total possible points.

D. Interactive Roster – CRT only

1. Students who are DNP in a subject are reported with scaled score=200 and performance level='DNP'.
2. Students who are Incomplete in a subject are reported with their earned scaled score and performance level='INC' on the interactive roster.
3. Students who are first-year LEP and who complete the reading test are reported with their earned scaled score and performance level and are included in school, system and state level aggregations for all subjects unless otherwise excluded based on completeness in math or science.
4. Students who are first-year LEP and who do not complete the reading test are reported with their earned scaled score and performance level='LEP' for all subjects. These students are excluded from school, system and state level aggregations.
5. Students who participated in Alternate assessment are listed on the rosters. Their scaled score is blank and the performance level='ALT'. These students are not included in aggregations.
6. The items are reported using the released item number.
7. Students who took the Braille form are not included in any rawscore aggregations. These students have a scaleform other than 01.
8. The following students will have included set to 0 in tblscoreitem (these students are excluded from performance level aggregations):
 - a. The student did not participate in the subject (partstatus='F')
 - b. The student has partstatus='E'
 - c. The student is LEPfirst (LEPfirst='1' regardless of how many items attempted)
 - d. The student is incomplete in the subject.
 - e. The student took the alternate assessment (alt='1')
 - f. Student took the subject with nonstandard accommodations (NSA).
 - g. Student is NSAY (NSAY='1').
 - h. Student is NDAY (NDAY='1').
9. If the student took the Braille form (Braille='1'), included is set to 2. These students are excluded from raw score aggregations.
10. If students do not fall into any of the categories in numbers 8 and 9 above, included is set to '1'.
11. If partstatus='E' for any subject then interactive='0' otherwise interactive='1'. Students with interactive='0' are not available in the interactive site.
12. State level item averages do not include students with school type PRAS, PRNAS or SNE.

13. District level item averages do not include students who are marked nday='1'.
14. Only students whose partstatus is not 'E' for any subject are included in tblStuLongitudinal.
15. The filter column in tblItemAveragesLookup is the concatenation of the gender,ethnic,iep,lep,econdis,migrant and plan504 fields in that order.
16. RepType='0' for all records in tblItemAverages.

E. Summary Report

1. Section I (Distribution of Scores)
 - a. Distribution of Scores will be suppressed and left blank for systems/schools with N less than 10.
2. Section II (Subtest Results) Students with scaleform other than 01 are not included in Subtest Results.
 - a. Subtest Results will be suppressed and left blank for systems/schools with N less than 10.
 - b. A footnote reading "Results are suppressed when less than ten (10) students were assessed." will appear at the bottom of the first page of the report.
 - c. (Alt only) If the number of total possible points is less than 5 for any Standard, place a dash ("—") in the school, system, and state cells for that standard. A footnote will appear below this section reading "—There were too few score points to report on this standard, or no items on the test measured this standard."
3. Section III (Results for Subgroups of Students)
 - a. Performance level results for subgroups with N less than 10 are suppressed, and the footnote "* Less than 10 students were assessed." will appear. N is always reported.
 - b. CRT only: Count of students who are considered NSA for that subject excluding those students who are incomplete, nsay (at school level), nday (at school and system level) or FXS or SNE or PSNE or First year LEP or alt (general assessment report).
 - c. Count of First year LEP students excludes those students who are nsay (at school level), nday (at school or system level) or incomplete or FXS or SNE or PSNE or NSA or alt (general assessment).

V. Data File Rules

1. The following students are not included in the state file:
 - a. Alternate Assessment students (in CRT)
 - b. Homeschooled students (SNE)
 - c. Student is in school less than 180 hours (PSNE)
2. If the student receives a performance level 'LEP' on the student report in Reading, the student receives LEP for the Reading performance level in the state files.

3. Alt students who are halted are marked '1' in the halted field for that subject.
4. Students who take the Braille form of the test are flagged Braille='1' in the state and system level files.
5. In the system and school level files only the released scored items are included.
6. The following students are not included in the system level files:
 - a. Alternate Assessment students (in CRT)
 - b. Foreign Exchange students (FXS='1')
 - c. Homeschooled students (SNE)
 - d. Student is in school less than 180 hours (PSNE)
7. The following students are not included in the previous year school level files:
 - a. Alternate Assessment students (in CRT)
 - b. Foreign Exchange students (FXS='1')
 - c. Homeschooled students (SNE)
 - d. Student is in school less than 180 hours (PSNE)
8. (Alt only) Standard data are not suppressed based on the number of total possible points.

VI. Shipping Product Code Summary

1. School (ReportFor='1')

Grade	Report Name	ReportType	Subject	ContentCode	Quantity
04	Student Labels (CRT)	03	Science	00	1 set for each school
08	Student Labels (CRT)	03	Science	00	1 set for each school
10	Student Labels (CRT)	03	Science	00	1 set for each school
04	Student Report (CRT)	02	Reading Math and Science	00	1 for each student
08	Student Report (CRT)	02	Reading Math and Science	00	1 for each student

Grade	Report Name	ReportType	Subject	ContentCode	Quantity
10	Student Report (CRT)	02	Reading Math and Science	00	1 for each student
03	Student Labels (CRT-Alt)	07	Reading and Math	00	1 set for each school
04	Student Labels (CRT-Alt)	07	Reading, Math and Science	00	1 set for each school
05	Student Labels (CRT-Alt)	07	Reading and Math	00	1 set for each school
06	Student Labels (CRT-Alt)	07	Reading and Math	00	1 set for each school
07	Student Labels (CRT-Alt)	07	Reading and Math	00	1 set for each school
08	Student Labels (CRT-Alt)	07	Reading Math and Science	00	1 set for each school
10	Student Labels (CRT-Alt)	07	Reading Math and Science	00	1 set for each school
03	Student Report (CRT-Alt)	08	Reading and Math	00	1 for each student

Grade	Report Name	ReportType	Subject	ContentCode	Quantity
04	Student Report (CRT-Alt)	08	Reading Math and Science	00	1 for each student
05	Student Report (CRT-Alt)	08	Reading Math	00	1 for each student
06	Student Report (CRT-Alt)	08	Reading and Math	00	1 for each student
07	Student Report (CRT-Alt)	08	Reading and Math	00	1 for each student
08	Student Report (CRT-Alt)	08	Reading Math and Science	00	1 for each student
10	Student Report (CRT-Alt)	08	Reading Math and Science	00	1 for each student
00	Interp. Guide	04		00	1 per school

Appendix A

1. Items not available on the Braille form

Grade	Form	Content	Positon	IABS#	
4	FT	Science	6	237575	Omit
4	FT	Science	7	209675	Omit
4	FT	Science	29	209666	Omit
4	FT	Science	46	237571	Omit
4	Common	Science	53	120019	Omit
4	Common	Science	59	75517	Omit

Note: Braille students with an item that could not be administered on the Braille test – on the student report suppress the student’s raw score for content standards that contain the excluded item.

Data File Deliverables: Files Produced

- CRT State Level Data Files
 - Results Data File
 - All Grades combined
 - Layout: Studentdatafilelayout.xls
 - Filename: Studentdatafile.csv
 - Raw Data
 - All Grades combined
 - Layout: Rawdatalayout.xls
 - Filename: RawData.csv
 - Plus Data
 - All grades combined
 - Layout: Plusdatalayout.xls
 - Filename: Plusdata.csv
 - Scored Data
 - All grades combined
 - Layout: Scoreddatalayout.xls
 - Filename: Scoreddata.csv
 - Test Meta-Data
 - All grades combined
 - Layout: Testmetadatalayout.xls
 - TestMetaData.csv
- CRT – Alternate State Level Data File
 - Results Data File
 - All Grades combined
 - Layout: AltStateStudentDataFileLayout.xls
 - Filename: Altstudentdatafile.csv
- CRT System and School Slice Data files (no changes)
- CRT-Alternate System and School Slice Data files (no changes)

APPENDIX B—PARTICIPATION RATES

Table B-1. 2013–14 MontCAS: Summary of Participation by Demographic Category—Science

<i>Description</i>	<i>Number Tested</i>	<i>Percent Tested</i>
Special Education	3,159	10.08
Title 1	109	0.35
Low Income	12485	39.84
American Indian or Alaskan Native	3,794	12.11
Asian	336	1.07
Hispanic	1,219	3.89
Black or African American	406	1.3
White, Non-Hispanic	25,477	81.29
Native Hawaiian/Other Pacific Islander	107	0.34
Female	15,206	48.52
Male	16,133	51.48
Limited English Proficient	717	2.29
Migrant	86	0.27
Plan 504	428	1.37
All Students	31,340	100

APPENDIX C—ACCOMMODATION FREQUENCIES

Table C-1. 2013–14 MontCAS: Numbers of Students Tested With Accommodations by Accommodation Type and Grade—Science

<i>Accommodation Code</i>	<i>Grade 4</i>	<i>Grade 8</i>	<i>Grade 10</i>
SCIAccom01	127	49	22
SCIAccom02	234	126	135
SCIAccom04	111	45	48
SCIAccom05	788	570	406
SCIAccom06	231	77	35
SCIAccom07	554	244	191
SCIAccom08	504	205	133
SCIAccom09	18	3	0
SCIAccom10	4	1	5
SCIAccom12	2	2	1
SCIAccom13	1	0	0
SCIAccom14	2	0	0
SCIAccom15	1	0	0
SCIAccom16	2	5	1
SCIAccom17	0	0	0
SCIAccom18	8	1	0
SCIAccom19	175	36	17
SCIAccom20	10	0	0
SCIAccom21	1	0	0
SCIAccom22	716	304	216
SCIAccom23	9	5	1
SCIAccom24	53	18	28
SCIAccom25	65	41	19
SCIAccom26	2	1	1
SCIAccom27	3	0	0
SCIAccom28	1	2	0
SCIAccom33	0	0	0

Figure C-1. 2013–14 MontCAS: Accommodations—Science

<i>Accommodation</i>	<i>Description</i>
SCIAccom01	Change in Administration Time: Test is administered at a time of day or a day of the week based on student needs.
SCIAccom02	Session Duration: Test is administered in appropriate blocks of time for individual student needs, followed by rest breaks.
SCIAccom04	Individual Administration: Test is administered in a one-to-one situation.
SCIAccom05	Small Group Administration: Test is administered to a small group of students.
SCIAccom06	Reduce Distracters: Student is seated at a carrel or other physical arrangement that reduces visual distractions.
SCIAccom07	Alternative Setting: Test is administered to a student in a different setting.
SCIAccom08	Change in Personnel: Test is administered by other personnel known to the student (e.g., LEP, Title I, special education teacher).
SCIAccom09	Home Setting: Test is administered to the student by school personnel in their home.
SCIAccom10	Front Row Seating: Student is seated at the front of the classroom when taking the test.

continued

<i>Accommodation</i>	<i>Description</i>
SCIAccom12	Magnification: Student used equipment to magnify test materials.
SCIAccom13	Student (not groups of students) wears equipment to reduce environmental noises.
SCIAccom14	Template: Student uses a template. An example is a piece of card stock that has a window cut out that enables the student to focus by isolating lines of text or items.
SCIAccom15	Amplification: Student uses amplification equipment (e.g., hearing aid or auditory trainer) while taking test.
SCIAccom16	Writing Tools: Student uses a typewriter or word processor (without activating spell check).
SCIAccom17	Voice Activation: Student speaks response into computer equipped with voice-activation software.
SCIAccom18	Bilingual Dictionary: Student uses a bilingual dictionary.
SCIAccom19	Dictation: Student dictates answers to a test administrator who records them in the Answer Booklet.
SCIAccom20	Writing Tools: Student marks or writes answers with the assistance of a technology device or special equipment.
SCIAccom21	Assistive Technology: Another form of assistive technology routinely used by the student (that does not change intent or test content).
SCIAccom22	Oral Presentation: The test administrator must read the test items and answer choices word-for-word and in exactly the order as presented.
SCIAccom23	Test Interpretation: Tests, including directions, are interpreted for students who are deaf or hearing-impaired.
SCIAccom24	Test Directions with Verification: An administrator gives test directions with verification (by using a highlighter) so that student understands them.
SCIAccom25	Test Directions Support: An administrator assists student in understanding test directions, including giving directions in native language.
SCIAccom26	Braille: Braille version of the test was used by the student.
SCIAccom27	Large Print: A large-print version of the test is used by student.
SCIAccom28	Other: With verification from the OPI in advance of the testing window, some other approved accommodation is used by student.
SCIAccom33	Other: With verification from the OPI in advance of the testing window, some other approved accommodation is used by student.

APPENDIX D—ITEM REVIEW COMMITTEE MEMBERS

**Table D-1. 2013–14 Montana CRT: Bias Item Review Committee Members
April 15, 2013**

<i>Name</i>	<i>District</i>
Karen Hutchison	Kalispell
Ellen Parchen	Missoula
Heather Dunn	Eureka
Carol Shipley	Great Falls
Karen Polari	Sidney
Theresa Romo	Wolf Point
Nina Miller	Billings
Kathleen Gilboy	Carroll College
Robin Hehn	Columbus
Robyn Ross	Billings
Christy Jobman	Huntley

**Table D-2. 2013–14 Montana CRT: Benchmarking Committee Members
May 7–8, 2014**

<i>Name</i>	<i>Content</i>
Nina Miller	Science
Paul Tackes	Science

**Table D-3. 2013–14 Montana CRT: Item Statistical Review Committee Members
June 24–25, 2013**

<i>Name</i>	<i>Position</i>
Katie Capp	Science Teacher
Amanda Becker	Science Teacher
Catherine Blee	Science Teacher
Allyson Hoof	Teacher
Marie Judisch	4 th Grade Teacher
Connie Warner	Math Specialist
Kellen Alger	Teacher
Kris Gardner	6-8 th Grade Science

APPENDIX E—ITEM-LEVEL CLASSICAL STATISTICS

**Table E-1. 2013–14 MontCAS: Item-Level Classical Test Theory Statistics—
Science Grade 4**

Item:		Difficulty	Discrimination	Percent Omitted	Item:		Difficulty	Discrimination	Percent Omitted
Number	Type				Number	Type			
120003	MC	0.69	0.40	0	134687	MC	0.72	0.31	0
168075	MC	0.88	0.36	0	159622	MC	0.71	0.30	0
120304	MC	0.61	0.37	0	75889	MC	0.49	0.43	0
237525	MC	0.51	0.27	0	56422	MC	0.63	0.31	0
209689	MC	0.82	0.31	0	75421	MC	0.51	0.28	1
120310	MC	0.81	0.34	0	60127	MC	0.50	0.28	1
159631	MC	0.67	0.45	0	39231	MC	0.95	0.17	0
75729	MC	0.59	0.24	0	120560	MC	0.77	0.19	0
39336	MC	0.39	0.39	0	209647	MC	0.84	0.29	0
209682	MC	0.82	0.34	0	60171	MC	0.58	0.45	0
159604	MC	0.63	0.23	0	120019	MC	0.47	0.12	0
120545	MC	0.73	0.35	0	120000	MC	0.82	0.32	0
208902	MC	0.81	0.30	0	75510	MC	0.88	0.37	0
134841	MC	0.85	0.31	0	166239	MC	0.68	0.24	0
209600	MC	0.51	0.27	0	134675	MC	0.68	0.46	0
75522	MC	0.78	0.37	0	159607	MC	0.41	0.28	0
237628	MC	0.72	0.34	1	75517	MC	0.62	0.31	0
257333	CR	0.41	0.35	1	56001	MC	0.47	0.38	0
159633	MC	0.49	0.37	0	209660	MC	0.58	0.30	0
75912	MC	0.73	0.32	0	166247	MC	0.55	0.41	0
75831	MC	0.87	0.40	0	57863	MC	0.61	0.32	0
134736	MC	0.72	0.45	0	75408	MC	0.53	0.25	1
75502	MC	0.44	0.31	0	75884	MC	0.56	0.36	0
209656	MC	0.58	0.25	0	166779	MC	0.68	0.41	1
209597	MC	0.71	0.33	0	159638	CR	0.26	0.51	1
75692	MC	0.53	0.31	0					
134754	MC	0.45	0.29	0					
120077	MC	0.84	0.23	0					
209686	MC	0.80	0.34	0					
166258	MC	0.80	0.33	0					

**Table E-2. 2013–14 MontCAS: Item-Level Classical Test Theory Statistics—
Science Grade 8**

Item:		Difficulty	Discrimination	Percent Omitted	Item:		Difficulty	Discrimination	Percent Omitted
Number	Type				Number	Type			
89693	MC	0.59	0.30	0	210198	MC	0.60	0.26	0
237689	MC	0.68	0.29	0	237617	MC	0.83	0.31	0
158515	MC	0.68	0.42	0	89895	MC	0.84	0.31	0
210221	MC	0.34	0.10	0	237688	MC	0.55	0.14	0
237651	MC	0.41	0.31	0	210173	MC	0.50	0.35	0
56846	MC	0.74	0.25	0	39659	MC	0.50	0.28	0
125949	MC	0.78	0.35	0	237540	MC	0.57	0.12	0
210209	MC	0.49	0.29	0	158487	CR	0.45	0.53	1
89382	MC	0.73	0.29	0	89420	MC	0.83	0.28	0
121184	MC	0.69	0.40	0					

continued

Item:		Difficulty	Discrimination	Percent Omitted
Number	Type			
210337	MC	0.70	0.15	0
75906	MC	0.68	0.29	0
237508	MC	0.68	0.42	0
210232	MC	0.55	0.21	0
89630	MC	0.44	0.21	0
210217	MC	0.58	0.35	0
122027	MC	0.74	0.33	0
54130	MC	0.58	0.38	0
212779	MC	0.58	0.39	0
122019	MC	0.59	0.27	0
158528	MC	0.58	0.37	0
158464	MC	0.60	0.30	0
122742	MC	0.84	0.34	0
89888	MC	0.67	0.37	0
237523	MC	0.62	0.44	0
158569	MC	0.47	0.29	0
39587	MC	0.64	0.19	1
89752	MC	0.68	0.29	0

Item:		Difficulty	Discrimination	Percent Omitted
Number	Type			
237518	MC	0.70	0.23	0
158491	MC	0.52	0.28	0
39745	MC	0.74	0.43	0
89691	MC	0.57	0.42	0
39782	MC	0.57	0.44	0
122029	MC	0.64	0.33	0
237610	MC	0.58	0.32	0
210215	MC	0.73	0.35	0
237506	MC	0.72	0.39	0
56773	MC	0.80	0.34	0
89426	MC	0.62	0.35	0
237692	MC	0.64	0.32	0
89800	MC	0.59	0.43	0
158476	MC	0.51	0.28	0
210216	MC	0.69	0.44	1
210244	MC	0.46	0.29	0
158578	MC	0.31	0.28	1
121235	CR	0.29	0.43	1

**Table E-3. 2013–14 MontCAS: Item-Level Classical Test Theory Statistics—
Science Grade 10**

Item:		Difficulty	Discrimination	Percent Omitted
Number	Type			
119654	MC	0.77	0.13	0
237818	MC	0.71	0.38	0
75447	MC	0.53	0.24	0
75876	MC	0.53	0.31	0
120026	MC	0.70	0.34	0
209055	MC	0.46	0.22	0
209061	MC	0.64	0.44	0
75802	MC	0.73	0.38	0
75445	MC	0.40	0.25	0
158423	MC	0.51	0.30	0
158625	MC	0.43	0.32	0
158433	MC	0.58	0.47	0
55209	MC	0.37	0.24	0
119893	MC	0.40	0.28	0
75968	MC	0.74	0.50	0
52946	MC	0.92	0.29	0
159445	MC	0.41	0.14	0
52993	CR	0.31	0.61	6
52276	MC	0.74	0.29	0
159436	MC	0.58	0.14	0
119797	MC	0.74	0.27	0
161153	MC	0.86	0.39	0
158443	MC	0.35	0.17	0
158449	MC	0.81	0.45	0
134799	MC	0.83	0.44	0

Item:		Difficulty	Discrimination	Percent Omitted
Number	Type			
53750	MC	0.71	0.43	0
75442	MC	0.53	0.37	0
75735	MC	0.53	0.13	0
75764	MC	0.85	0.43	0
119674	MC	0.60	0.22	0
120064	MC	0.81	0.32	0
158435	MC	0.35	0.20	0
134560	MC	0.50	0.23	0
237736	MC	0.60	0.34	1
75435	MC	0.78	0.38	0
75970	MC	0.59	0.45	0
75852	MC	0.85	0.37	0
158428	MC	0.88	0.37	0
134539	MC	0.60	0.32	1
75433	MC	0.49	0.40	1
206952	MC	0.58	0.29	1
119945	MC	0.81	0.33	1
134497	MC	0.36	0.13	1
75799	MC	0.62	0.29	1
75749	MC	0.43	0.27	1
75863	MC	0.65	0.49	1
206906	MC	0.45	0.37	1
75456	MC	0.57	0.29	1
130591	MC	0.59	0.43	1

continued

Item:		Difficulty	Discrimination	Percent Omitted
Number	Type			
119952	MC	0.80	0.45	1
206890	MC	0.40	0.34	1
134541	MC	0.49	0.31	1

Item:		Difficulty	Discrimination	Percent Omitted
Number	Type			
159493	MC	0.62	0.37	1
134471	MC	0.73	0.45	1
56042	CR	0.33	0.52	3

APPENDIX F—ITEM-LEVEL SCORE-POINT DISTRIBUTIONS

**Table F-1. 2013–14 MontCAS: Item-Level Score-Point Distributions for Constructed-Response Items—
Science**

<i>Grade</i>	<i>Item Number</i>	<i>Total Possible Points</i>	<i>Percent of Students at Score Points</i>				
			<i>0</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
4	257333	4	15.48	31.48	32.15	10.70	9.33
	159638	4	38.34	27.79	24.16	6.33	2.44
8	158487	4	28.42	13.98	18.62	20.95	16.78
	121235	4	34.28	27.73	25.02	7.77	4.07
10	52993	4	37.46	24.21	7.65	14.42	10.04
	56042	4	10.21	54.14	22.82	8.35	1.77

APPENDIX G—NUMBER OF ITEMS CLASSIFIED INTO DIFFERENTIAL ITEM FUNCTIONING CATEGORIES

Table G-1. 2013–14 MontCAS: Number of Items Classified as “Low” or “High” DIF, Overall and by Group Favored—Science

Grade	Item Type	Group		Number of Items	Number “Low”			Number “High”		
		Reference	Focal		Total	Favoring		Total	Favoring	
						Reference	Focal		Reference	Focal
4	MC	Male	Female	53	9	7	2	1	0	1
		White	Hispanic	53	3	3	0	0	0	0
			Native American	53	4	4	0	0	0	0
		No Disability	Disability	53	4	2	2	1	1	0
		Not Low Income	Low Income	53	0	0	0	0	0	0
		Not Limited English Proficient	Limited English Proficient	53	10	9	1	8	6	2
	OR	Male	Female	2	0	0	0	0	0	0
		White	Hispanic	2	0	0	0	0	0	0
			Native American	2	0	0	0	0	0	0
		No Disability	Disability	2	0	0	0	0	0	0
Not Low Income		Low Income	2	0	0	0	0	0	0	
	Not Limited English Proficient	Limited English Proficient	2	1	1	0	0	0	0	
8	MC	Male	Female	53	10	6	4	0	0	0
		White	Hispanic	53	4	3	1	0	0	0
			Native American	53	3	3	0	0	0	0
		No Disability	Disability	53	9	8	1	3	3	0
		Not Low Income	Low Income	53	0	0	0	0	0	0
		Not Limited English Proficient	Limited English Proficient	53	9	8	1	16	13	3
	OR	Male	Female	2	0	0	0	0	0	0
		White	Hispanic	2	0	0	0	0	0	0
			Native American	2	0	0	0	0	0	0
		No Disability	Disability	2	1	1	0	0	0	0
Not Low Income		Low Income	2	0	0	0	0	0	0	
	Not Limited English Proficient	Limited English Proficient	2	0	0	0	2	2	0	

continued

Grade	Item Type	Group		Number of Items	Number "Low"			Number "High"			
		Reference	Focal		Total	Favoring		Total	Favoring		
						Reference	Focal		Reference	Focal	
10	MC	Male	Female	53	10	6	4	2	2	0	
		White	Hispanic		53	3	2	1	0	0	0
			Native American		53	3	2	1	0	0	0
		No Disability	Disability		53	12	9	3	2	2	0
		Not Low Income	Low Income		53	0	0	0	0	0	0
		Not Limited English Proficient	Limited English Proficient		53	0	0	0	0	0	0
	OR	Male	Female		2	0	0	0	0	0	0
		White	Hispanic		2	0	0	0	0	0	0
			Native American		2	1	1	0	0	0	0
		No Disability	Disability		2	2	2	0	0	0	0
		Not Low Income	Low Income		2	0	0	0	0	0	0
		Not Limited English Proficient	Limited English Proficient		2	0	0	0	0	0	0

APPENDIX H—ITEM RESPONSE THEORY CALIBRATION RESULTS

**Table H-1. 2013–14 MontCAS: IRT Parameters for Dichotomous Items—
Science Grade 4**

<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>c</i>	<i>SE (c)</i>	<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>c</i>	<i>SE (c)</i>
120003	1.00000	0.00000	-0.36305	0.00000	0.00000	0.00000	134687	1.00000	0.00000	-0.70843	0.00000	0.00000	0.00000
168075	1.00000	0.00000	-1.13502	0.00000	0.00000	0.00000	159622	1.00000	0.00000	-0.53620	0.00000	0.00000	0.00000
120304	1.00000	0.00000	-0.11908	0.00000	0.00000	0.00000	75889	1.00000	0.00000	0.07412	0.00000	0.00000	0.00000
237525	1.00000	0.00000	0.01956	0.00000	0.00000	0.00000	56422	1.00000	0.00000	-0.23316	0.00000	0.00000	0.00000
209689	1.00000	0.00000	-0.87266	0.00000	0.00000	0.00000	75421	1.00000	0.00000	0.07731	0.01206	0.00000	0.00000
120310	1.00000	0.00000	-0.59108	0.00000	0.00000	0.00000	60127	1.00000	0.00000	0.09948	0.00000	0.00000	0.00000
159631	1.00000	0.00000	-0.38479	0.00000	0.00000	0.00000	39231	1.00000	0.00000	-1.81296	0.02649	0.00000	0.00000
75729	1.00000	0.00000	-0.11884	0.00000	0.00000	0.00000	120560	1.00000	0.00000	-0.74194	0.00000	0.00000	0.00000
39336	1.00000	0.00000	0.37079	0.00000	0.00000	0.00000	209647	1.00000	0.00000	-0.93169	0.00000	0.00000	0.00000
209682	1.00000	0.00000	-0.75966	0.00000	0.00000	0.00000	60171	1.00000	0.00000	-0.05913	0.00000	0.00000	0.00000
159604	1.00000	0.00000	-0.32890	0.00000	0.00000	0.00000	120019	1.00000	0.00000	0.06248	0.00000	0.00000	0.00000
120545	1.00000	0.00000	-0.56669	0.00000	0.00000	0.00000	120000	1.00000	0.00000	-0.80385	0.00000	0.00000	0.00000
208902	1.00000	0.00000	-0.72359	0.00000	0.00000	0.00000	75510	1.00000	0.00000	-1.12652	0.00000	0.00000	0.00000
134841	1.00000	0.00000	-0.91606	0.00000	0.00000	0.00000	166239	1.00000	0.00000	-0.58806	0.00000	0.00000	0.00000
209600	1.00000	0.00000	0.04454	0.00000	0.00000	0.00000	134675	1.00000	0.00000	-0.37925	0.00000	0.00000	0.00000
75522	1.00000	0.00000	-0.60102	0.00000	0.00000	0.00000	159607	1.00000	0.00000	0.36635	0.00000	0.00000	0.00000
237628	1.00000	0.00000	-0.54463	0.00000	0.00000	0.00000	75517	1.00000	0.00000	-0.23689	0.01241	0.00000	0.00000
159633	1.00000	0.00000	0.11688	0.01206	0.00000	0.00000	56001	1.00000	0.00000	0.01426	0.00000	0.00000	0.00000
75912	1.00000	0.00000	-0.55770	0.01341	0.00000	0.00000	209660	1.00000	0.00000	-0.14561	0.00000	0.00000	0.00000
75831	1.00000	0.00000	-1.15076	0.01736	0.00000	0.00000	166247	1.00000	0.00000	-0.08012	0.00000	0.00000	0.00000
134736	1.00000	0.00000	-0.54719	0.00000	0.00000	0.00000	57863	1.00000	0.00000	-0.17730	0.00000	0.00000	0.00000
75502	1.00000	0.00000	0.11290	0.00000	0.00000	0.00000	75408	1.00000	0.00000	0.02399	0.00000	0.00000	0.00000
209656	1.00000	0.00000	-0.04479	0.00000	0.00000	0.00000	75884	1.00000	0.00000	-0.14381	0.00000	0.00000	0.00000
209597	1.00000	0.00000	-0.52203	0.00000	0.00000	0.00000	166779	1.00000	0.00000	-0.39387	0.01282	0.00000	0.00000
75692	1.00000	0.00000	-0.15205	0.00000	0.00000	0.00000							
134754	1.00000	0.00000	0.32130	0.00000	0.00000	0.00000							
120077	1.00000	0.00000	-0.76385	0.00000	0.00000	0.00000							
209686	1.00000	0.00000	-0.82312	0.00000	0.00000	0.00000							
166258	1.00000	0.00000	-0.91240	0.00000	0.00000	0.00000							

**Table H-2. 2013–14 MontCAS: IRT Parameters for Polytomous Items—
Science Grade 4**

<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>D0</i>	<i>SE (D0)</i>	<i>D1</i>	<i>SE (D1)</i>	<i>D2</i>	<i>SE (D2)</i>	<i>D3</i>	<i>SE (D3)</i>	<i>D4</i>	<i>SE (D4)</i>
159638	1	0	0.8192	0	0	0	0.67083	0	0.56812	0	-0.47397	0	-0.76499	0
257333	1	0	0.26615	0.0061	0	0	0.84241	0.01781	0.20474	0.01469	-0.70039	0.02033	-0.34676	0.02576

**Table H-3. 2013–14 MontCAS: IRT Parameters for Dichotomous Items—
Science Grade 8**

<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>c</i>	<i>SE (c)</i>	<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>c</i>	<i>SE (c)</i>
89693	1.00000	0.00000	-0.29967	0.00000	0.00000	0.00000	54130	1.00000	0.00000	-0.27684	0.00000	0.00000	0.00000
237689	1.00000	0.00000	-0.22786	0.00000	0.00000	0.00000	212779	1.00000	0.00000	0.02064	0.00000	0.00000	0.00000
158515	1.00000	0.00000	-0.22766	0.00000	0.00000	0.00000	122019	1.00000	0.00000	-0.15087	0.00000	0.00000	0.00000
210221	1.00000	0.00000	0.57427	0.01286	0.00000	0.00000	158528	1.00000	0.00000	-0.10879	0.00000	0.00000	0.00000
237651	1.00000	0.00000	0.43113	0.00000	0.00000	0.00000	158464	1.00000	0.00000	-0.15625	0.00000	0.00000	0.00000
56846	1.00000	0.00000	-0.40150	0.00000	0.00000	0.00000	122742	1.00000	0.00000	-0.81105	0.00000	0.00000	0.00000
125949	1.00000	0.00000	-0.64829	0.00000	0.00000	0.00000	89888	1.00000	0.00000	-0.32813	0.00000	0.00000	0.00000
210209	1.00000	0.00000	0.22338	0.00000	0.00000	0.00000	237523	1.00000	0.00000	-0.18390	0.00000	0.00000	0.00000
89382	1.00000	0.00000	-0.53717	0.01374	0.00000	0.00000	158569	1.00000	0.00000	0.13872	0.00000	0.00000	0.00000
121184	1.00000	0.00000	-0.22855	0.00000	0.00000	0.00000	39587	1.00000	0.00000	-0.27819	0.00000	0.00000	0.00000
210198	1.00000	0.00000	-0.15055	0.00000	0.00000	0.00000	89752	1.00000	0.00000	-0.41017	0.00000	0.00000	0.00000
237617	1.00000	0.00000	-0.80804	0.00000	0.00000	0.00000	237518	1.00000	0.00000	-0.38845	0.00000	0.00000	0.00000
89895	1.00000	0.00000	-0.96498	0.00000	0.00000	0.00000	158491	1.00000	0.00000	-0.04452	0.00000	0.00000	0.00000
237688	1.00000	0.00000	0.05454	0.00000	0.00000	0.00000	39745	1.00000	0.00000	-0.41132	0.00000	0.00000	0.00000
210173	1.00000	0.00000	0.17349	0.00000	0.00000	0.00000	89691	1.00000	0.00000	-0.04377	0.01235	0.00000	0.00000
39659	1.00000	0.00000	0.13753	0.01225	0.00000	0.00000	39782	1.00000	0.00000	0.00803	0.00000	0.00000	0.00000
237540	1.00000	0.00000	-0.09343	0.00000	0.00000	0.00000	122029	1.00000	0.00000	-0.37172	0.00000	0.00000	0.00000
89420	1.00000	0.00000	-0.74537	0.00000	0.00000	0.00000	237610	1.00000	0.00000	-0.16478	0.00000	0.00000	0.00000
210337	1.00000	0.00000	-0.30159	0.00000	0.00000	0.00000	210215	1.00000	0.00000	-0.53012	0.01371	0.00000	0.00000
75906	1.00000	0.00000	-0.35446	0.01303	0.00000	0.00000	237506	1.00000	0.00000	-0.22184	0.00000	0.00000	0.00000
237508	1.00000	0.00000	-0.17792	0.00000	0.00000	0.00000	56773	1.00000	0.00000	-0.74439	0.00000	0.00000	0.00000
210232	1.00000	0.00000	-0.13840	0.00000	0.00000	0.00000	89426	1.00000	0.00000	-0.32411	0.00000	0.00000	0.00000
89630	1.00000	0.00000	0.25087	0.00000	0.00000	0.00000	237692	1.00000	0.00000	-0.22400	0.00000	0.00000	0.00000
210217	1.00000	0.00000	-0.03148	0.00000	0.00000	0.00000	89800	1.00000	0.00000	-0.15591	0.00000	0.00000	0.00000
122027	1.00000	0.00000	-0.54596	0.00000	0.00000	0.00000							

continued

<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>c</i>	<i>SE (c)</i>
158476	1.00000	0.00000	-0.06047	0.00000	0.00000	0.00000
210216	1.00000	0.00000	-0.43934	0.00000	0.00000	0.00000

<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>c</i>	<i>SE (c)</i>
210244	1.00000	0.00000	0.30214	0.00000	0.00000	0.00000
158578	1.00000	0.00000	0.60762	0.00000	0.00000	0.00000

**Table H-4. 2013–14 MontCAS: IRT Parameters for Polytomous Items—
Science Grade 8**

<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>D0</i>	<i>SE (D0)</i>	<i>D1</i>	<i>SE (D1)</i>	<i>D2</i>	<i>SE (D2)</i>	<i>D3</i>	<i>SE (D3)</i>	<i>D4</i>	<i>SE (D4)</i>
121235	1	0	0.56018	0	0	0	0.45017	0	0.28482	0	-0.48524	0	-0.24975	0
158487	1	0	0.20771	0	0	0	-0.09783	0	0.39952	0	0.02819	0	-0.32988	0

**Table H-5. 2013–14 MontCAS: IRT Parameters for Dichotomous Items—
Science Grade 10**

<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>c</i>	<i>SE (c)</i>	<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>c</i>	<i>SE (c)</i>
119654	1.00000	0.00000	-0.47387	0.00000	0.00000	0.00000	120064	1.00000	0.00000	-0.83395	0.00000	0.00000	0.00000
237818	1.00000	0.00000	-0.41544	0.00000	0.00000	0.00000	158435	1.00000	0.00000	0.59952	0.00000	0.00000	0.00000
75447	1.00000	0.00000	0.05889	0.00000	0.00000	0.00000	134560	1.00000	0.00000	0.10775	0.00000	0.00000	0.00000
75876	1.00000	0.00000	-0.27268	0.00000	0.00000	0.00000	237736	1.00000	0.00000	-0.14633	0.00000	0.00000	0.00000
120026	1.00000	0.00000	-0.45243	0.00000	0.00000	0.00000	75435	1.00000	0.00000	-0.79741	0.00000	0.00000	0.00000
209055	1.00000	0.00000	0.27864	0.00000	0.00000	0.00000	75970	1.00000	0.00000	-0.13565	0.00000	0.00000	0.00000
209061	1.00000	0.00000	-0.15065	0.00000	0.00000	0.00000	75852	1.00000	0.00000	-0.93178	0.00000	0.00000	0.00000
75802	1.00000	0.00000	-0.42214	0.00000	0.00000	0.00000	158428	1.00000	0.00000	-1.04931	0.00000	0.00000	0.00000
75445	1.00000	0.00000	0.41013	0.00000	0.00000	0.00000	134539	1.00000	0.00000	0.02252	0.00000	0.00000	0.00000
158423	1.00000	0.00000	0.12818	0.00000	0.00000	0.00000	75433	1.00000	0.00000	-0.05999	0.00000	0.00000	0.00000
158625	1.00000	0.00000	0.14057	0.00000	0.00000	0.00000	206952	1.00000	0.00000	-0.13147	0.00000	0.00000	0.00000
158433	1.00000	0.00000	0.01638	0.00000	0.00000	0.00000	119945	1.00000	0.00000	-1.08698	0.00000	0.00000	0.00000
55209	1.00000	0.00000	0.44951	0.00000	0.00000	0.00000	134497	1.00000	0.00000	0.26710	0.00000	0.00000	0.00000
119893	1.00000	0.00000	0.29366	0.00000	0.00000	0.00000	75799	1.00000	0.00000	-0.41260	0.00000	0.00000	0.00000
75968	1.00000	0.00000	-0.54787	0.00000	0.00000	0.00000	75749	1.00000	0.00000	0.34585	0.00000	0.00000	0.00000
52946	1.00000	0.00000	-1.22780	0.00000	0.00000	0.00000	75863	1.00000	0.00000	-0.41040	0.00000	0.00000	0.00000
159445	1.00000	0.00000	0.30625	0.00000	0.00000	0.00000	206906	1.00000	0.00000	0.24904	0.00000	0.00000	0.00000
52276	1.00000	0.00000	-0.69099	0.00000	0.00000	0.00000	75456	1.00000	0.00000	-0.11919	0.00000	0.00000	0.00000
159436	1.00000	0.00000	0.04233	0.00000	0.00000	0.00000	130591	1.00000	0.00000	-0.15455	0.00000	0.00000	0.00000
119797	1.00000	0.00000	-0.46102	0.00000	0.00000	0.00000	119952	1.00000	0.00000	-1.05526	0.00000	0.00000	0.00000
161153	1.00000	0.00000	-0.88790	0.00000	0.00000	0.00000	206890	1.00000	0.00000	0.23789	0.00000	0.00000	0.00000
158443	1.00000	0.00000	0.40086	0.00000	0.00000	0.00000	134541	1.00000	0.00000	0.01129	0.00000	0.00000	0.00000
158449	1.00000	0.00000	-0.73226	0.00000	0.00000	0.00000	159493	1.00000	0.00000	-0.24229	0.00000	0.00000	0.00000
134799	1.00000	0.00000	-0.80975	0.00000	0.00000	0.00000	134471	1.00000	0.00000	-0.81545	0.00000	0.00000	0.00000
53750	1.00000	0.00000	-0.40420	0.00000	0.00000	0.00000							
75442	1.00000	0.00000	-0.06837	0.00000	0.00000	0.00000							
75735	1.00000	0.00000	-0.07617	0.00000	0.00000	0.00000							
75764	1.00000	0.00000	-0.97266	0.00000	0.00000	0.00000							
119674	1.00000	0.00000	-0.14317	0.00000	0.00000	0.00000							

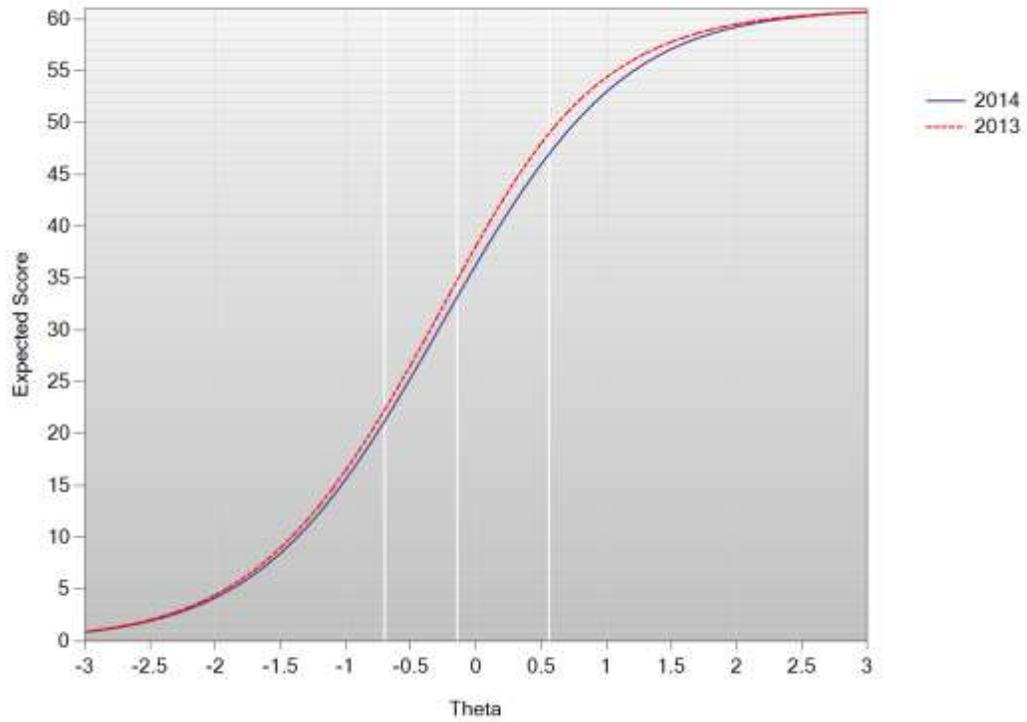
**Table H-6. 2013–14 MontCAS: IRT Parameters for Polytomous Items—
Science Grade 10**

<i>IREF</i>	<i>a</i>	<i>SE (a)</i>	<i>b</i>	<i>SE (b)</i>	<i>D0</i>	<i>SE (D0)</i>	<i>D1</i>	<i>SE (D1)</i>	<i>D2</i>	<i>SE (D2)</i>	<i>D3</i>	<i>SE (D3)</i>	<i>D4</i>	<i>SE (D4)</i>
52993	1	0	0.68767	0	0	0	0.43798	0	-0.2428	0	0.85209	0	-1.04727	0
56042	1	0	0.28061	0	0	0	1.36823	0	-0.06351	0	-0.45494	0	-0.84978	0

APPENDIX I—TEST CHARACTERISTIC CURVES AND TEST INFORMATION FUNCTIONS

Figure I-1. 2013–14 MontCAS: Science Grade 4 Plots
Top: Test Characteristic Curve Bottom: Test Information Function

Test Characteristic Curve: Science Grade 4



Test Information Function: Science Grade 4

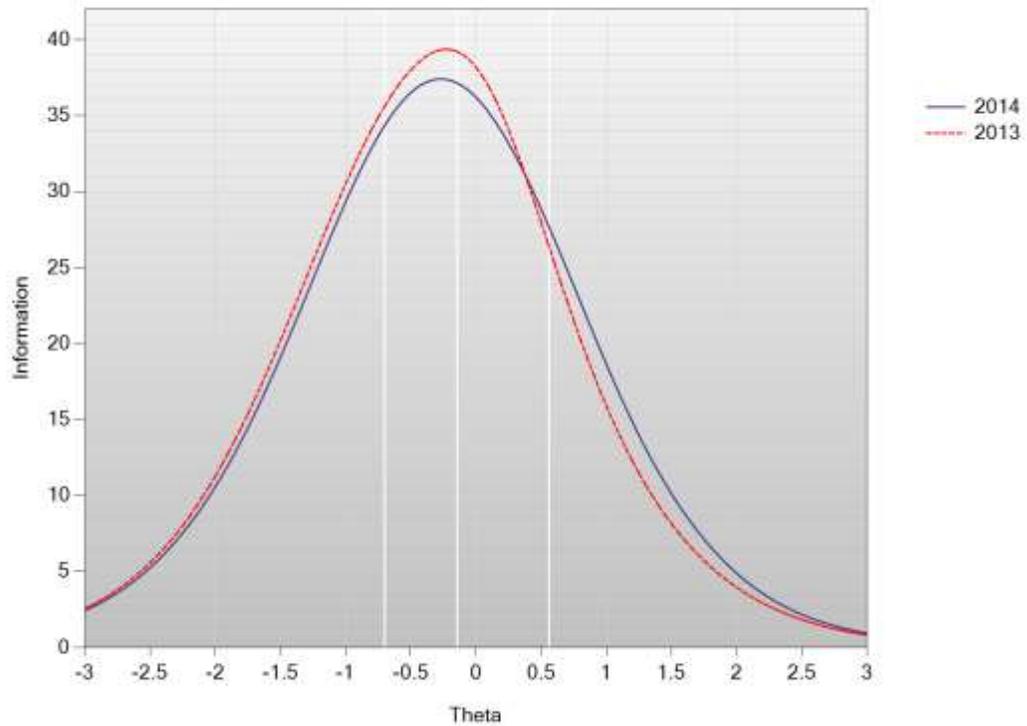


Figure I-2. 2013–14 MontCAS: Science Grade 8 Plots
Top: Test Characteristic Curve Bottom: Test Information Function

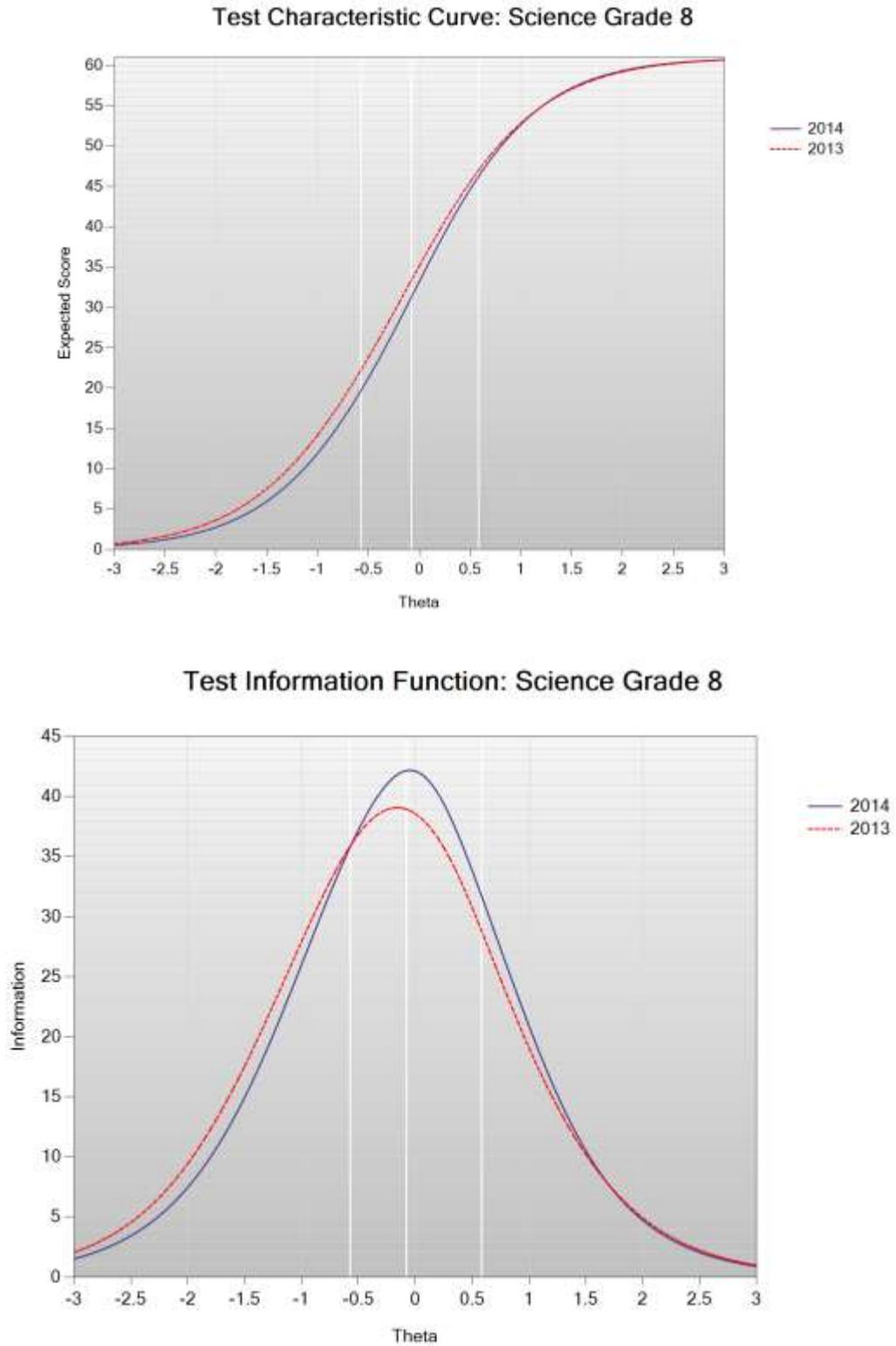
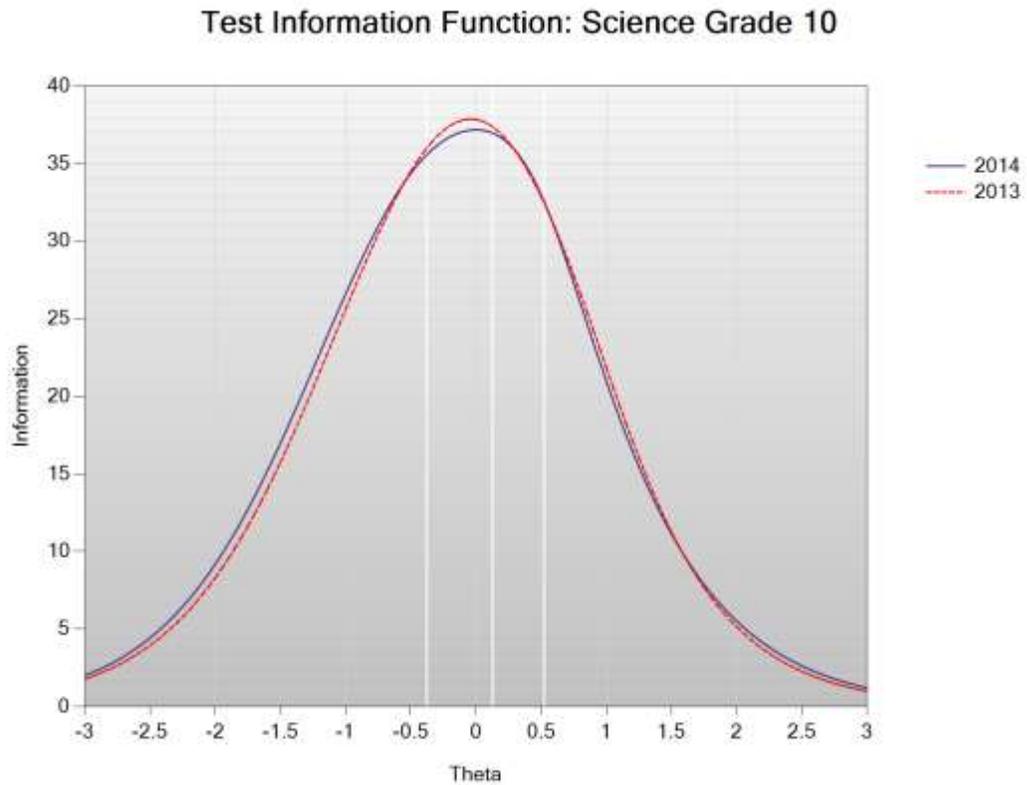
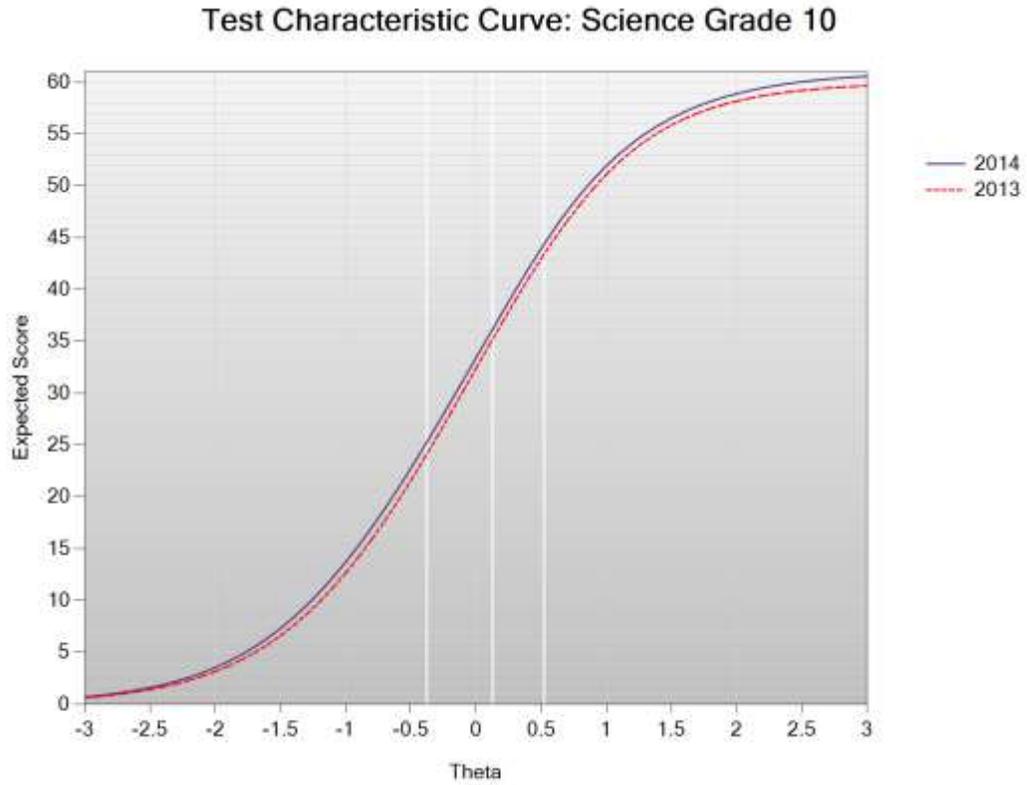


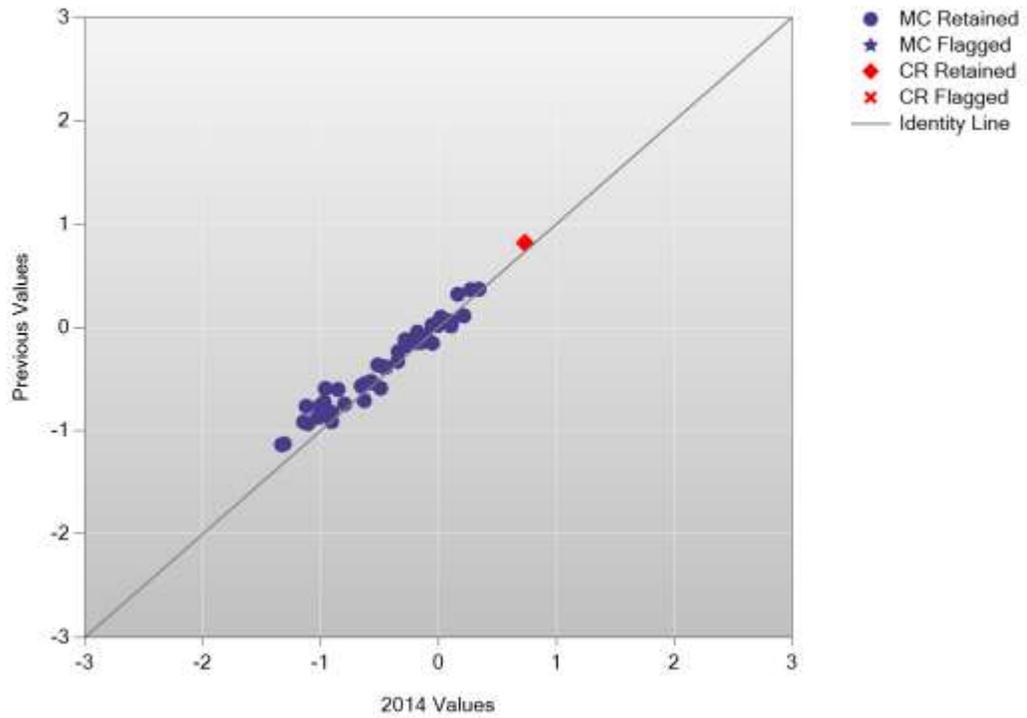
Figure I-3. 2013–14 MontCAS: Science Grade 10 Plots
Top: Test Characteristic Curve Bottom: Test Information Function



APPENDIX J—*b*-PLOTS

Figure J-1. 2013–14 MontCAS: *b*-Plots
Top: Science Grade 4 Bottom: Science Grade 8

B/B Plot: Science Grade 4



B/B Plot: Science Grade 8

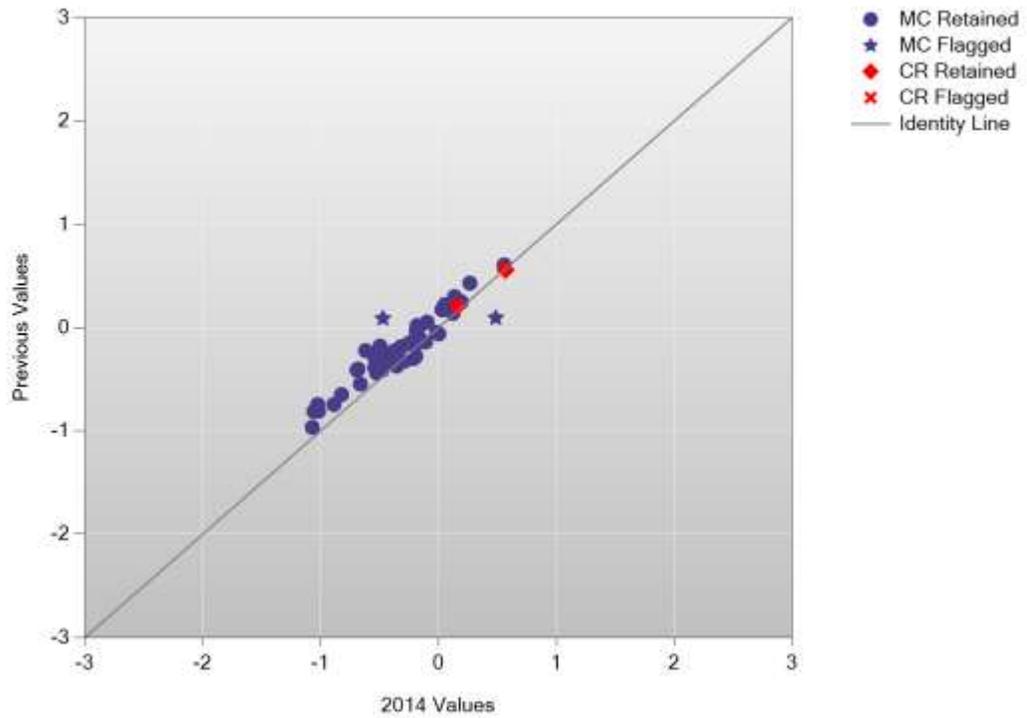
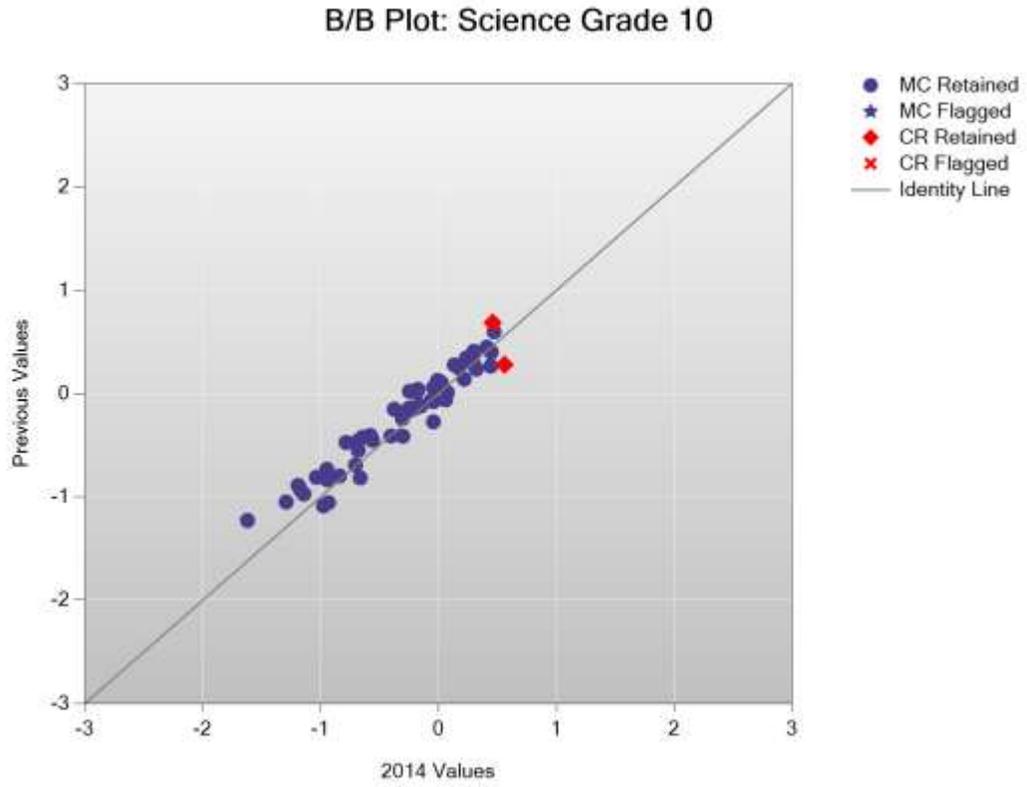
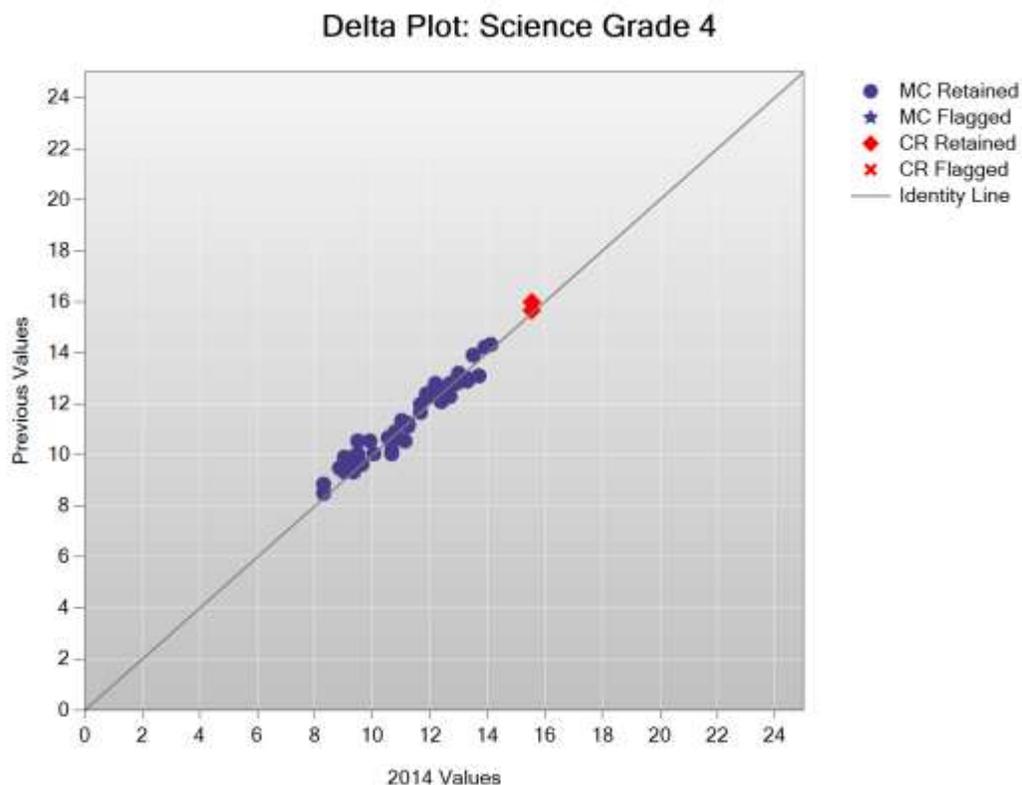


Figure J-2. 2013–14 MontCAS: *b*-Plot
Science Grade 10



**APPENDIX K—ANALYSES OF EQUATING ITEMS (DELTA AND
RESCORE ANALYSES)**

**Figure K-1. 2013–14 MontCAS: Delta Plot—
Science Grade 4**



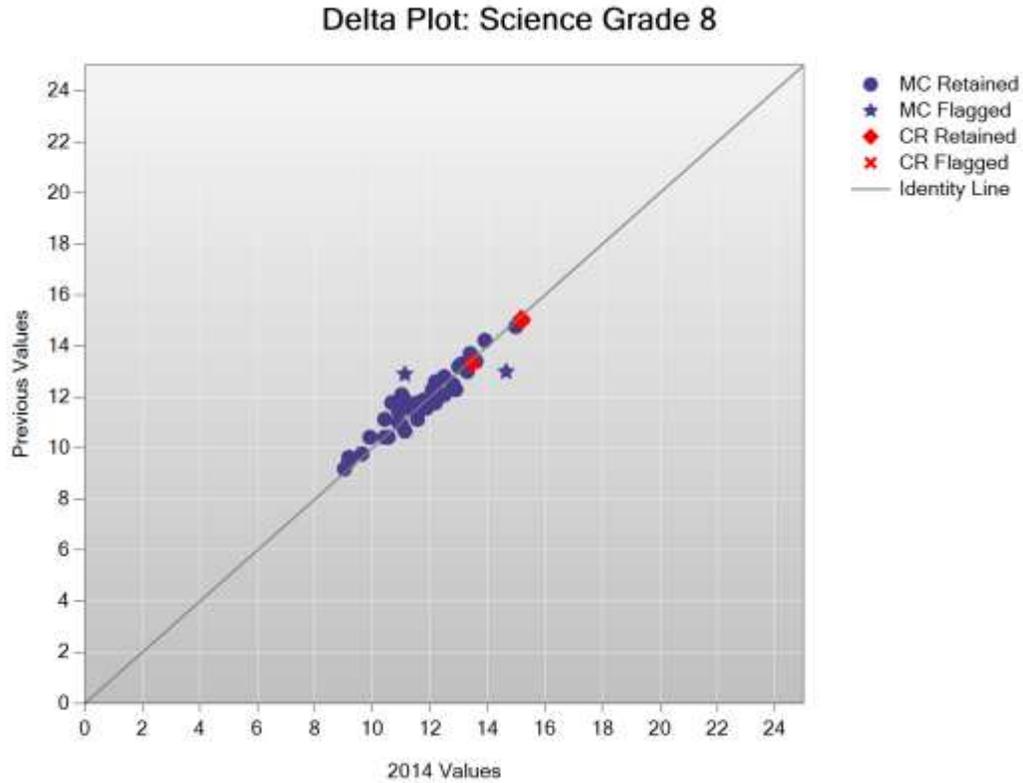
**Table K-2. 2013–14 MontCAS: Delta Analysis Results—
Science Grade 4**

<i>IREF</i>	<i>Mean</i>		<i>Delta</i>		<i>Maximum</i>	<i>Discard</i>	<i>Standardized Difference</i>
	<i>Old</i>	<i>New</i>	<i>Old</i>	<i>New</i>			
120000	0.79000	0.82000	9.77432	9.33854	1	False	-0.41464
120003	0.66000	0.69000	11.35015	11.01660	1	False	-0.57725
120019	0.50000	0.47000	13.00000	13.30108	1	False	0.33034
120077	0.78000	0.84000	9.91123	9.02217	1	False	1.70063
120304	0.56000	0.61000	12.39612	11.88272	1	False	0.45611
120310	0.73000	0.81000	10.54875	9.48841	1	False	2.61454
120545	0.72000	0.73000	10.66863	10.54875	1	False	-1.15298
120560	0.77000	0.77000	10.04461	10.04461	1	False	-0.47849
134675	0.68000	0.67000	11.12920	11.24035	1	False	-0.17866
134687	0.77000	0.72000	10.04461	10.66863	1	False	2.39666
134736	0.76000	0.72000	10.17479	10.66863	1	False	1.77140
134754	0.41000	0.45000	13.91018	13.50265	1	False	0.26465
134841	0.81000	0.85000	9.48841	8.85427	1	False	0.44340
159604	0.63000	0.63000	11.67259	11.67259	1	False	-0.79709
159607	0.38000	0.41000	14.22192	13.91018	1	False	-0.11570
159622	0.70000	0.71000	10.90240	10.78646	1	False	-1.18054
159631	0.67000	0.67000	11.24035	11.24035	1	False	-0.71250
159638	0.25250	0.26250	15.66657	15.54263	4	False	-0.69824

continued

<i>IREF</i>	<i>Mean</i>		<i>Delta</i>		<i>Maximum</i>	<i>Discard</i>	<i>Standardized Difference</i>
	<i>Old</i>	<i>New</i>	<i>Old</i>	<i>New</i>			
159638	0.22750	0.26250	15.98842	15.54263	4	False	0.84765
166239	0.73000	0.68000	10.54875	11.12920	1	False	2.09728
166247	0.54000	0.55000	12.59827	12.49735	1	False	-1.40486
166258	0.80000	0.80000	9.63352	9.63352	1	False	-0.39803
168075	0.85000	0.88000	8.85427	8.30005	1	False	-0.04900
208902	0.77000	0.81000	10.04461	9.48841	1	False	0.19310
209597	0.72000	0.71000	10.66863	10.78646	1	False	-0.05773
209600	0.51000	0.51000	12.89972	12.89972	1	False	-1.03725
209647	0.82000	0.84000	9.33854	9.02217	1	False	-1.05008
209656	0.55000	0.57000	12.49735	12.29450	1	False	-0.95492
209660	0.57000	0.58000	12.29450	12.19243	1	False	-1.38912
209682	0.78000	0.82000	9.91123	9.33854	1	False	0.24297
209686	0.80000	0.80000	9.63352	9.63352	1	False	-0.39803
209689	0.82000	0.82000	9.33854	9.33854	1	False	-0.34031
237525	0.52000	0.51000	12.79939	12.89972	1	False	-0.55531
237628	0.73000	0.72000	10.54875	10.66863	1	False	-0.02478
39336	0.37000	0.39000	14.32741	14.11728	1	False	-0.56320
56001	0.51000	0.47000	12.89972	13.30108	1	False	0.81198
56422	0.60000	0.63000	11.98661	11.67259	1	False	-0.54265
57863	0.57000	0.61000	12.29450	11.88272	1	False	-0.03199
60127	0.48000	0.50000	13.20061	13.00000	1	False	-0.82759
60171	0.52000	0.58000	12.79939	12.19243	1	False	0.96610
75408	0.52000	0.53000	12.79939	12.69892	1	False	-1.36755
75502	0.49000	0.43000	13.10028	13.70550	1	False	1.71203
75510	0.87000	0.88000	8.49444	8.30005	1	False	-1.07072
75522	0.73000	0.78000	10.54875	9.91123	1	False	0.66645
75692	0.57000	0.53000	12.29450	12.69892	1	False	0.94453
75729	0.56000	0.59000	12.39612	12.08982	1	False	-0.49808
75729	0.55000	0.59000	12.49735	12.08982	1	False	-0.01185
75884	0.59000	0.56000	12.08982	12.39612	1	False	0.53253
75889	0.51000	0.49000	12.89972	13.10028	1	False	-0.11322

**Figure K-2. 2013–14 MontCAS: Delta Plot—
Science Grade 8**



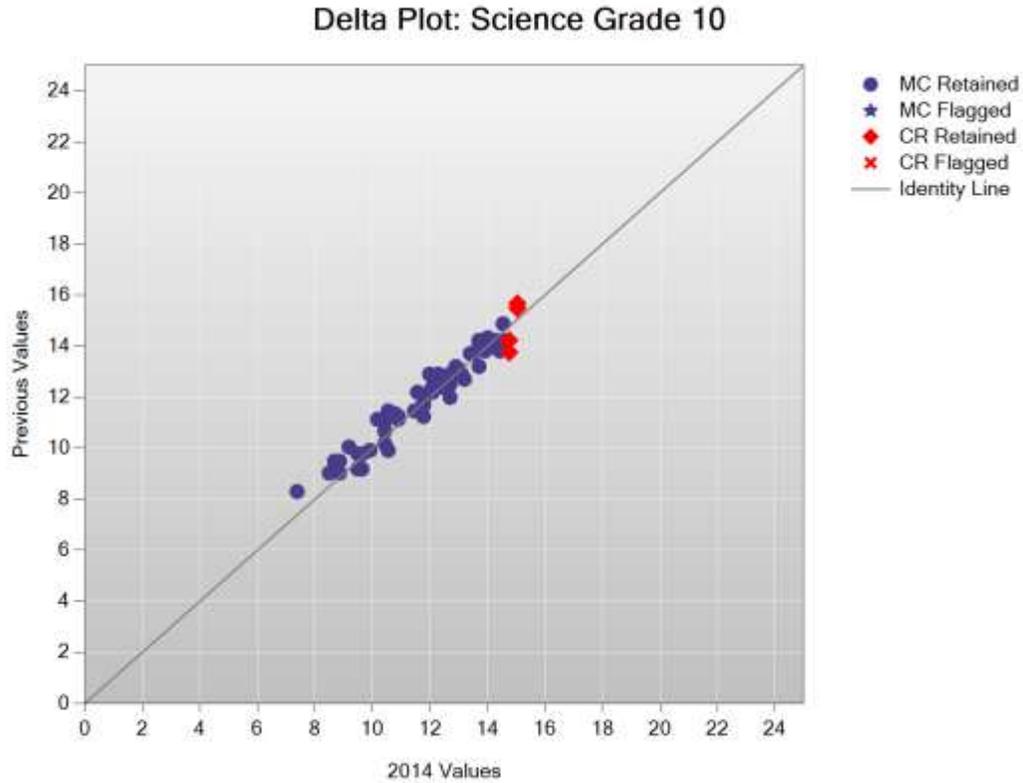
**Table K-3. 2013–14 MontCAS: Delta Analysis Results—
Science Grade 8**

IREF	Mean		Delta		Maximum	Discard	Standardized Difference
	Old	New	Old	New			
121184	0.59000	0.69000	12.08982	11.01660	1	False	1.84177
121235	0.30250	0.29250	15.06889	15.18438	4	False	-0.56430
121235	0.31000	0.29250	14.98340	15.18438	4	False	-0.84715
122019	0.61000	0.59000	11.88272	12.08982	1	False	-0.17862
122027	0.74000	0.73000	10.42662	10.54875	1	False	0.10478
122029	0.68000	0.64000	11.12920	11.56616	1	False	0.77368
122742	0.80000	0.83000	9.63352	9.18334	1	False	-0.89106
125949	0.74000	0.78000	10.42662	9.91123	1	False	-0.40867
158464	0.60000	0.60000	11.98661	11.98661	1	False	-0.82598
158476	0.57000	0.51000	12.29450	12.89972	1	False	0.84187
158487	0.46750	0.45250	13.32622	13.47739	4	False	-0.87193
158487	0.46750	0.45250	13.32622	13.47739	4	False	-0.87193
158491	0.55000	0.52000	12.49735	12.79939	1	False	-0.12445
158515	0.62000	0.68000	11.77808	11.12920	1	False	0.47909
158528	0.61000	0.58000	11.88272	12.19243	1	False	0.12326
158569	0.50000	0.47000	13.00000	13.30108	1	False	-0.31137
158578	0.33000	0.31000	14.75965	14.98340	1	False	-0.99609
210173	0.48000	0.50000	13.20061	13.00000	1	False	-0.31865

continued

<i>IREF</i>	<i>Mean</i>		<i>Delta</i>		<i>Maximum</i>	<i>Discard</i>	<i>Standardized Difference</i>
	<i>Old</i>	<i>New</i>	<i>Old</i>	<i>New</i>			
210198	0.61000	0.60000	11.88272	11.98661	1	False	-0.48227
210209	0.47000	0.49000	13.30108	13.10028	1	False	-0.28129
210216	0.70000	0.69000	10.90240	11.01660	1	False	-0.09283
210217	0.54000	0.58000	12.59827	12.19243	1	False	0.06450
210221	0.50000	0.34000	13.00000	14.64985	1	True	3.65687
210232	0.59000	0.55000	12.08982	12.49735	1	False	0.33523
210244	0.43000	0.46000	13.70550	13.40173	1	False	0.16976
210337	0.65000	0.70000	11.45872	10.90240	1	False	0.08981
212779	0.54000	0.58000	12.59827	12.19243	1	False	0.06450
237506	0.62000	0.72000	11.77808	10.66863	1	False	1.83414
237508	0.61000	0.68000	11.88272	11.12920	1	False	0.82531
237518	0.69000	0.70000	11.01660	10.90240	1	False	-0.80665
237523	0.61000	0.62000	11.88272	11.77808	1	False	-1.08375
237540	0.58000	0.57000	12.19243	12.29450	1	False	-0.60104
237610	0.60000	0.58000	11.98661	12.19243	1	False	-0.22045
237617	0.80000	0.83000	9.63352	9.18334	1	False	-0.89106
237651	0.38000	0.41000	14.22192	13.91018	1	False	0.38241
237688	0.52000	0.55000	12.79939	12.49735	1	False	-0.16724
237689	0.62000	0.68000	11.77808	11.12920	1	False	0.47909
237692	0.62000	0.64000	11.77808	11.56616	1	False	-0.80649
39587	0.64000	0.64000	11.56616	11.56616	1	False	-0.67196
39745	0.74000	0.74000	10.42662	10.42662	1	False	-0.25454
39782	0.59000	0.57000	12.08982	12.29450	1	False	-0.26158
54130	0.61000	0.58000	11.88272	12.19243	1	False	0.12326
54130	0.62000	0.58000	11.77808	12.19243	1	False	0.46947
56773	0.79000	0.80000	9.77432	9.63352	1	False	-0.42985
56846	0.68000	0.74000	11.12920	10.42662	1	False	0.39944
75906	0.51000	0.68000	12.89972	11.12920	1	True	4.18997
89420	0.81000	0.83000	9.48841	9.18334	1	False	-0.80844
89426	0.64000	0.61000	11.56616	11.88272	1	False	0.25939
89630	0.46000	0.44000	13.40173	13.60388	1	False	-0.74962
89693	0.61000	0.59000	11.88272	12.08982	1	False	-0.17862
89752	0.72000	0.68000	10.66863	11.12920	1	False	1.01186
89800	0.59000	0.59000	12.08982	12.08982	1	False	-0.86378
89800	0.57000	0.59000	12.29450	12.08982	1	False	-0.63859
89888	0.64000	0.67000	11.56616	11.24035	1	False	-0.54900
89888	0.63000	0.67000	11.67259	11.24035	1	False	-0.19691
89895	0.83000	0.84000	9.18334	9.02217	1	False	-0.27331

**Figure K-3. 2013–14 MontCAS: Delta Plot—
Science Grade 10**



**Table K-4. 2013–14 MontCAS: Delta Analysis Results—
Science Grade 10**

IREF	Mean		Delta		Maximum	Discard	Standardized Difference
	Old	New	Old	New			
119654	0.68000	0.76000	11.12920	10.17479	1	False	1.36618
119674	0.58000	0.60000	12.19243	11.98661	1	False	-1.19287
119797	0.69000	0.74000	11.01660	10.42662	1	False	-0.04699
119893	0.42000	0.41000	13.80757	13.91018	1	False	-0.88360
119945	0.83000	0.81000	9.18334	9.48841	1	False	1.11297
119952	0.83000	0.80000	9.18334	9.63352	1	False	1.66372
120026	0.68000	0.70000	11.12920	10.90240	1	False	-1.39556
120064	0.79000	0.80000	9.77432	9.63352	1	False	-0.73635
130591	0.58000	0.59000	12.19243	12.08982	1	False	-1.23357
134471	0.78000	0.73000	9.91123	10.54875	1	False	2.18150
134497	0.42000	0.36000	13.80757	14.43384	1	False	1.10400
134539	0.51000	0.60000	12.89972	11.98661	1	False	1.67961
134541	0.53000	0.48000	12.69892	13.20061	1	False	0.92562
134560	0.49000	0.50000	13.10028	13.00000	1	False	-1.35236
134799	0.77000	0.83000	10.04461	9.18334	1	False	0.72461
158423	0.48000	0.51000	13.20061	12.89972	1	False	-0.56425
158428	0.84000	0.87000	9.02217	8.49444	1	False	-0.81293
158433	0.53000	0.58000	12.69892	12.19243	1	False	0.08291

continued

<i>IREF</i>	<i>Mean</i>		<i>Delta</i>		<i>Maximum</i>	<i>Discard</i>	<i>Standardized Difference</i>
	<i>Old</i>	<i>New</i>	<i>Old</i>	<i>New</i>			
158435	0.32000	0.35000	14.87080	14.54128	1	False	-0.01205
158443	0.38000	0.35000	14.22192	14.54128	1	False	-0.17093
158449	0.79000	0.81000	9.77432	9.48841	1	False	-1.28710
158625	0.48000	0.43000	13.20061	13.70550	1	False	0.80448
159436	0.51000	0.57000	12.89972	12.29450	1	False	0.51097
159445	0.41000	0.41000	13.91018	13.91018	1	False	-1.30031
159493	0.62000	0.62000	11.77808	11.77808	1	False	-0.73408
161153	0.81000	0.86000	9.48841	8.67872	1	False	0.38111
206890	0.41000	0.40000	13.91018	14.01339	1	False	-0.90857
206906	0.43000	0.45000	13.70550	13.50265	1	False	-0.80229
206952	0.53000	0.58000	12.69892	12.19243	1	False	0.08291
209055	0.43000	0.46000	13.70550	13.40173	1	False	-0.41927
209061	0.58000	0.64000	12.19243	11.56616	1	False	0.40299
237736	0.58000	0.60000	12.19243	11.98661	1	False	-1.19287
237818	0.67000	0.70000	11.24035	10.90240	1	False	-0.94418
52276	0.76000	0.74000	10.17479	10.42662	1	False	0.64756
52946	0.88000	0.92000	8.30005	7.37971	1	False	0.48549
52993	0.25250	0.30500	15.66657	15.04029	4	False	1.32570
52993	0.26750	0.30500	15.48157	15.04029	4	False	0.57435
53750	0.66000	0.71000	11.35015	10.78646	1	False	-0.05821
55209	0.38000	0.37000	14.22192	14.32741	1	False	-0.98270
56042	0.38000	0.33000	14.22192	14.75965	4	False	0.65792
56042	0.42500	0.33000	13.75647	14.75965	4	False	2.54821
75433	0.51000	0.49000	12.89972	13.10028	1	False	-0.27074
75435	0.78000	0.78000	9.91123	9.91123	1	False	-0.23829
75442	0.54000	0.53000	12.59827	12.69892	1	False	-0.56985
75445	0.37000	0.40000	14.32741	14.01339	1	False	-0.21515
75447	0.51000	0.53000	12.89972	12.69892	1	False	-1.02405
75456	0.56000	0.56000	12.39612	12.39612	1	False	-0.89821
75735	0.53000	0.53000	12.69892	12.69892	1	False	-0.97863
75735	0.55000	0.53000	12.49735	12.69892	1	False	-0.16003
75749	0.38000	0.43000	14.22192	13.70550	1	False	0.52508
75749	0.39000	0.43000	14.11728	13.70550	1	False	0.10008
75764	0.84000	0.85000	9.02217	8.85427	1	False	-0.63947
75799	0.63000	0.62000	11.67259	11.77808	1	False	-0.30566
75799	0.67000	0.62000	11.24035	11.77808	1	False	1.44975
75802	0.66000	0.73000	11.35015	10.54875	1	False	0.84406
75802	0.65000	0.73000	11.45872	10.54875	1	False	1.28499
75852	0.81000	0.85000	9.48841	8.85427	1	False	-0.28519
75863	0.65000	0.65000	11.45872	11.45872	1	False	-0.64926
75876	0.60000	0.53000	11.98661	12.69892	1	False	1.91420
75968	0.72000	0.74000	10.66863	10.42662	1	False	-1.35804
75970	0.56000	0.59000	12.39612	12.08982	1	False	-0.75736

**Table K-5. 2013–14 MontCAS: Rescore Analysis Results—
Science**

<i>Grade</i>	<i>IREF</i>	<i>Maximum</i>	<i>Mean</i>		<i>Standard Deviation</i>		<i>Effect Size</i>	<i>Discard</i>
			<i>Old</i>	<i>New</i>	<i>Old</i>	<i>New</i>		
4	257333	4	1.44608	1.51471	1.13260	1.26910	0.06059	False
8	158487	4	1.91176	1.94118	1.46958	1.42686	0.02001	False
	121235	4	1.17561	1.27805	1.14120	1.14864	0.08976	False
10	52993	4	1.13235	1.06863	1.25832	1.32991	-0.05064	False
	56042	4	1.80882	1.54412	1.06799	0.94322	-0.24785	False

APPENDIX L—SCORE DISTRIBUTIONS

**Table L-1. 2013–14 MontCAS: Performance Level Distributions—
Science**

Grade	Performance Level	Percent in Level		
		2013–14	2012–13	2011–12
4	4	20.65	18.37	14.11
	3	46.95	51.82	53.91
	2	26.18	23.99	26.38
	1	6.21	5.82	5.59
8	4	16.55	17.46	19.99
	3	50.55	47.62	46.71
	2	25.49	23.89	25.18
	1	7.42	11.04	8.12
10	4	19.5	21.03	21.56
	3	26.68	24.67	24.27
	2	35.07	32.06	32.6
	1	18.75	22.24	21.57

**Figure L-1. 2013–14 MontCAS: Scaled Score Percentages—
Science Grade 4**

Cumulative Scale Score Distributions: Science Grade 4

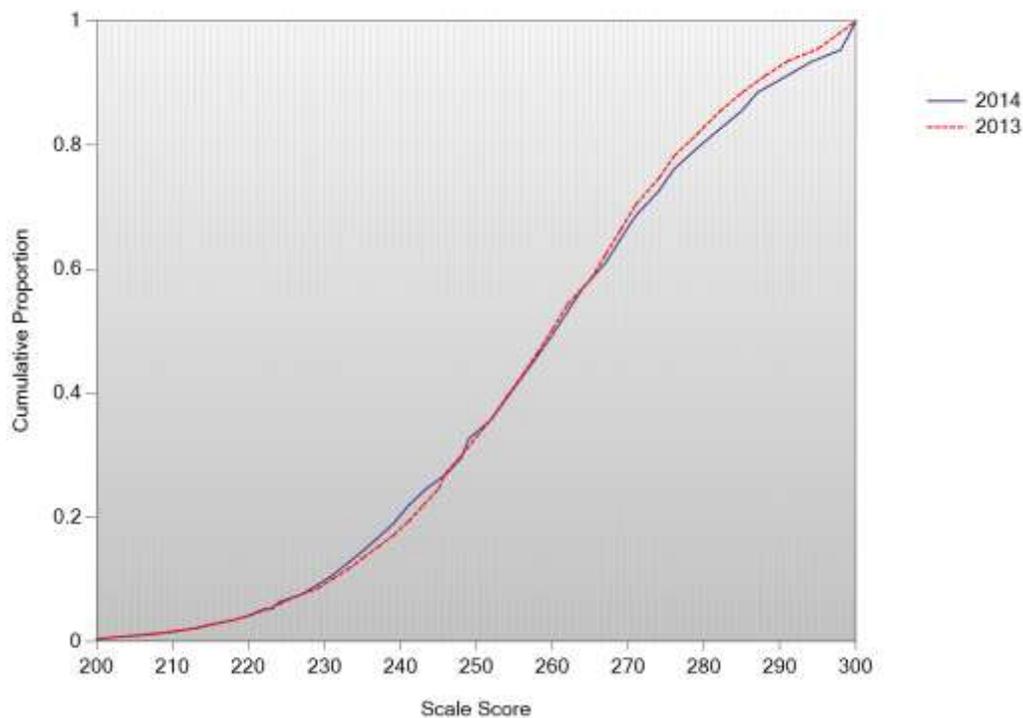
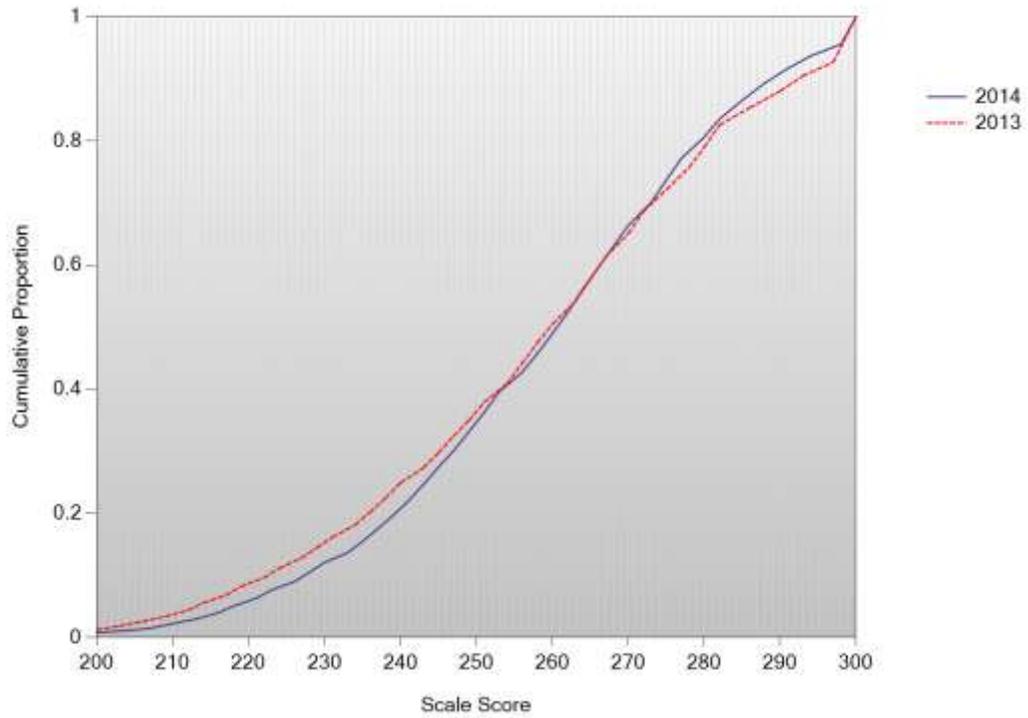
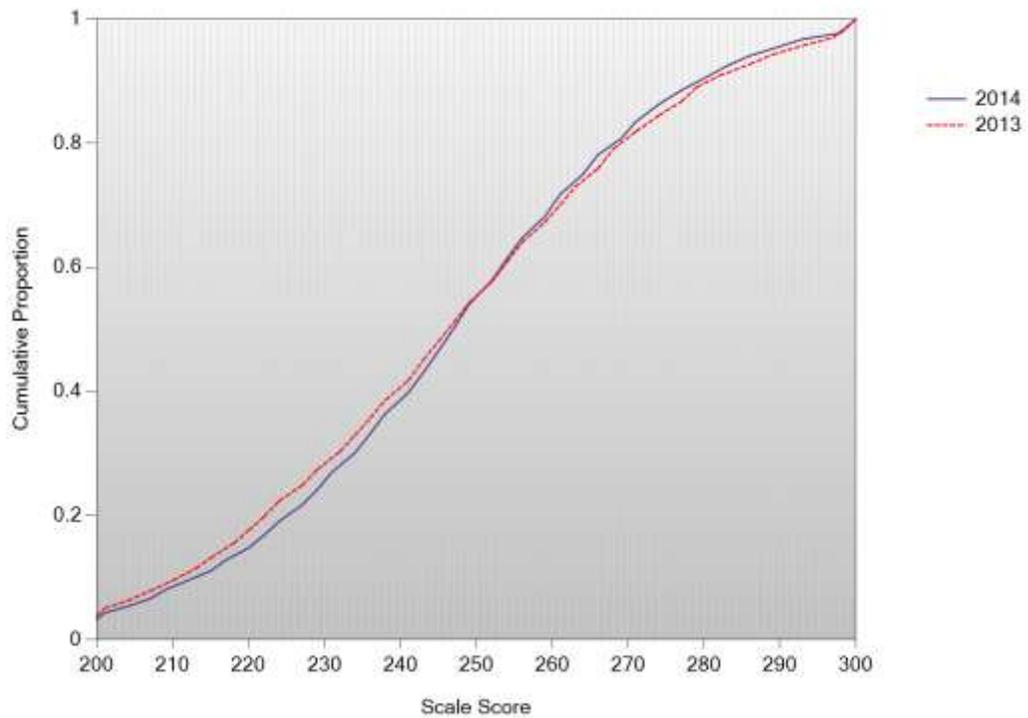


Figure L-2. 2013–14 MontCAS: Scaled Score Percentages
Top: Science Grade 8 Bottom: Science Grade 10

Cumulative Scale Score Distributions: Science Grade 8



Cumulative Scale Score Distributions: Science Grade 10



APPENDIX M—RAW TO SCALED SCORE LOOKUP TABLES

**Table M-1. 2013–14 MontCAS: Raw to Scaled Score Lookup Table—
Science Grade 4**

<i>Raw Score</i>	<i>2013–14</i>			<i>2012–13</i>		
	<i>Scaled Score</i>	<i>Standard Error</i>	<i>Performance Level</i>	<i>Scaled Score</i>	<i>Standard Error</i>	<i>Performance Level</i>
0	200	10.0	1	200	10.0	1
1	200	10.0	1	200	10.0	1
2	200	10.0	1	200	10.0	1
3	200	10.0	1	200	10.0	1
4	200	10.0	1	200	10.0	1
5	200	10.0	1	200	10.0	1
6	200	10.0	1	200	10.0	1
7	200	10.0	1	200	10.0	1
8	200	10.0	1	200	10.0	1
9	200	10.0	1	200	10.0	1
10	200	9.7	1	200	9.6	1
11	200	9.3	1	200	9.3	1
12	201	9.0	1	200	9.0	1
13	204	8.8	1	202	8.8	1
14	207	8.6	1	205	8.6	1
15	210	8.4	1	208	8.4	1
16	213	8.2	1	210	8.2	1
17	215	8.1	1	213	8.1	1
18	218	8.0	1	215	7.9	1
19	220	7.9	1	218	7.8	1
20	222	7.8	1	220	7.7	1
21	224	7.7	1	222	7.6	1
22	227	7.6	2	224	7.6	1
23	229	7.6	2	226	7.5	2
24	231	7.5	2	229	7.4	2
25	233	7.5	2	231	7.4	2
26	235	7.4	2	233	7.3	2
27	237	7.4	2	235	7.3	2
28	239	7.4	2	237	7.2	2
29	241	7.4	2	239	7.2	2
30	243	7.4	2	241	7.2	2
31	246	7.4	2	243	7.2	2
32	248	7.4	2	245	7.2	2
33	249	7.4	2	246	7.2	2
34	252	7.4	3	248	7.2	2
35	254	7.4	3	250	7.2	3
36	256	7.5	3	252	7.2	3
37	258	7.5	3	254	7.2	3
38	260	7.6	3	256	7.3	3
39	262	7.6	3	258	7.3	3
40	264	7.7	3	260	7.4	3
41	267	7.8	3	262	7.5	3
42	269	7.9	3	265	7.5	3
43	271	8.0	3	267	7.6	3
44	274	8.1	3	269	7.8	3

continued

Raw Score	2013–14			2012–13		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
45	276	8.2	3	271	7.9	3
46	279	8.4	3	274	8.1	3
47	282	8.5	4	276	8.3	3
48	285	8.7	4	279	8.5	3
49	287	9.0	4	282	8.8	4
50	291	9.3	4	285	9.1	4
51	294	9.6	4	288	9.4	4
52	298	10.0	4	291	9.8	4
53	300	10.0	4	295	10.0	4
54	300	10.0	4	300	10.0	4
55	300	10.0	4	300	10.0	4
56	300	10.0	4	300	10.0	4
57	300	10.0	4	300	10.0	4
58	300	10.0	4	300	10.0	4
59	300	10.0	4	300	10.0	4
60	300	10.0	4	300	10.0	4
61	300	10.0	4	300	10.0	4

**Table M-2. 2013–14 MontCAS: Raw to Scaled Score Lookup Table—
Science Grade 8**

Raw Score	2013–14			2012–13		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
0	200	10.0	1	200	10.0	1
1	200	10.0	1	200	10.0	1
2	200	10.0	1	200	10.0	1
3	200	10.0	1	200	10.0	1
4	200	10.0	1	200	10.0	1
5	200	10.0	1	200	10.0	1
6	200	10.0	1	200	10.0	1
7	200	10.0	1	200	10.0	1
8	200	10.0	1	200	10.0	1
9	200	10.0	1	200	10.0	1
10	200	10.0	1	200	10.0	1
11	200	10.0	1	200	10.0	1
12	204	9.9	1	200	10.0	1
13	207	9.6	1	200	9.8	1
14	210	9.4	1	203	9.6	1
15	213	9.1	1	206	9.4	1
16	216	9.0	1	209	9.2	1
17	218	8.8	1	212	9.0	1
18	221	8.6	1	214	8.9	1
19	223	8.5	1	217	8.8	1
20	226	8.4	2	219	8.6	1
21	228	8.3	2	222	8.5	1
22	230	8.2	2	224	8.4	1
23	233	8.1	2	227	8.4	2

continued

Raw Score	2013–14			2012–13		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
24	235	8.0	2	229	8.3	2
25	237	8.0	2	231	8.2	2
26	239	7.9	2	234	8.2	2
27	241	7.9	2	236	8.1	2
28	243	7.8	2	238	8.1	2
29	245	7.8	2	240	8.1	2
30	247	7.8	2	243	8.1	2
31	249	7.8	2	245	8.1	2
32	251	7.8	3	247	8.1	2
33	253	7.8	3	249	8.1	2
34	256	7.8	3	251	8.1	3
35	258	7.8	3	254	8.1	3
36	260	7.8	3	256	8.1	3
37	262	7.9	3	258	8.2	3
38	264	7.9	3	260	8.2	3
39	266	8.0	3	263	8.3	3
40	268	8.1	3	265	8.4	3
41	270	8.2	3	267	8.5	3
42	273	8.3	3	270	8.6	3
43	275	8.4	3	272	8.7	3
44	277	8.6	3	275	8.8	3
45	280	8.7	3	278	9.0	3
46	282	8.9	3	280	9.2	3
47	285	9.1	4	282	9.4	3
48	288	9.3	4	286	9.6	4
49	291	9.6	4	290	9.9	4
50	294	9.9	4	293	10.0	4
51	298	10.0	4	297	10.0	4
52	300	10.0	4	300	10.0	4
53	300	10.0	4	300	10.0	4
54	300	10.0	4	300	10.0	4
55	300	10.0	4	300	10.0	4
56	300	10.0	4	300	10.0	4
57	300	10.0	4	300	10.0	4
58	300	10.0	4	300	10.0	4
59	300	10.0	4	300	10.0	4
60	300	10.0	4	300	10.0	4
61	300	10.0	4	300	10.0	4

**Table M-3. 2013–14 MontCAS: Raw to Scaled Score Lookup Table—
Science Grade 10**

<i>Raw Score</i>	<i>2013–14</i>			<i>2012–13</i>		
	<i>Scaled Score</i>	<i>Standard Error</i>	<i>Performance Level</i>	<i>Scaled Score</i>	<i>Standard Error</i>	<i>Performance Level</i>
0	200	10.0	1	200	10.0	1
1	200	10.0	1	200	10.0	1
2	200	10.0	1	200	10.0	1
3	200	10.0	1	200	10.0	1
4	200	10.0	1	200	10.0	1
5	200	10.0	1	200	10.0	1
6	200	10.0	1	200	10.0	1
7	200	10.0	1	200	10.0	1
8	200	10.0	1	200	10.0	1
9	200	10.0	1	200	10.0	1
10	200	10.0	1	200	10.0	1
11	200	10.0	1	200	10.0	1
12	200	10.0	1	200	10.0	1
13	200	9.8	1	200	9.7	1
14	200	9.5	1	200	9.5	1
15	200	9.3	1	201	9.3	1
16	201	9.2	1	204	9.1	1
17	204	9.0	1	207	8.9	1
18	207	8.9	1	210	8.8	1
19	209	8.8	1	213	8.7	1
20	212	8.7	1	215	8.6	1
21	215	8.6	1	218	8.5	1
22	217	8.5	1	220	8.4	1
23	220	8.4	1	222	8.3	1
24	222	8.4	1	224	8.3	1
25	224	8.3	1	227	8.2	2
26	227	8.3	2	229	8.2	2
27	229	8.2	2	232	8.1	2
28	231	8.2	2	234	8.1	2
29	234	8.2	2	236	8.1	2
30	236	8.1	2	238	8.0	2
31	238	8.1	2	241	8.0	2
32	241	8.1	2	243	8.0	2
33	243	8.1	2	245	8.0	2
34	245	8.1	2	247	8.1	2
35	247	8.1	2	249	8.1	2
36	249	8.1	2	252	8.1	3
37	252	8.2	3	254	8.2	3
38	254	8.2	3	256	8.2	3
39	256	8.2	3	259	8.3	3
40	259	8.3	3	261	8.3	3
41	261	8.3	3	263	8.4	3
42	264	8.4	3	266	8.5	3
43	266	8.5	3	268	8.6	3
44	269	8.6	3	271	8.8	4

continued

<i>Raw Score</i>	<i>2013–14</i>			<i>2012–13</i>		
	<i>Scaled Score</i>	<i>Standard Error</i>	<i>Performance Level</i>	<i>Scaled Score</i>	<i>Standard Error</i>	<i>Performance Level</i>
45	271	8.8	4	274	8.9	4
46	274	8.9	4	277	9.1	4
47	277	9.1	4	279	9.3	4
48	280	9.4	4	282	9.5	4
49	283	9.6	4	286	9.8	4
50	286	10.0	4	289	10.0	4
51	290	10.0	4	293	10.0	4
52	293	10.0	4	297	10.0	4
53	298	10.0	4	300	10.0	4
54	300	10.0	4	300	10.0	4
55	300	10.0	4	300	10.0	4
56	300	10.0	4	300	10.0	4
57	300	10.0	4	300	10.0	4
58	300	10.0	4	300	10.0	4
59	300	10.0	4	300	10.0	4
60	300	10.0	4	300	10.0	4
61	300	10.0	4	-N/A	N/A	N/A

APPENDIX N—CLASSICAL RELIABILITY

**Table N-1. 2013–14 MontCAS: Subgroup Reliabilities—
Science**

Grade	Group	Number of Students	Raw Score			Alpha	SEM
			Maximum	Mean	Standard Deviation		
4	Special Education	1165	61	30.70	9.97	0.88	3.52
	Title 1	19	61	30.89	9.69	0.86	3.63
	Low Income	4924	61	34.41	9.77	0.87	3.48
	American Indian or Alaskan Native	1498	61	30.03	9.44	0.86	3.53
	Asian	130	61	39.03	9.84	0.88	3.38
	Hispanic	432	61	35.43	9.57	0.87	3.47
	Black or African American	157	61	34.03	9.42	0.86	3.51
	White, Non-Hispanic	8581	61	39.29	9.16	0.86	3.39
	Native Hawaiian/Other Pacific Islander	45	61	35.76	9.58	0.87	3.43
	Female	5327	61	37.50	9.71	0.87	3.45
	Male	5516	61	38.03	9.85	0.88	3.38
	Limited English Proficient	388	61	25.31	8.42	0.82	3.53
	Migrant	29	61	36.45	9.03	0.86	3.40
	Plan 504	82	61	38.54	10.33	0.89	3.35
	All Students	10844	61	37.77	9.78	0.88	3.42
8	Special Education	1017	61	25.80	9.51	0.85	3.67
	Title 1	24	61	29.71	7.18	0.72	3.80
	Low Income	4161	61	32.31	10.33	0.87	3.71
	American Indian or Alaskan Native	1238	61	28.04	10.07	0.86	3.72
	Asian	110	61	38.02	9.90	0.87	3.61
	Hispanic	403	61	32.34	9.91	0.86	3.74
	Black or African American	136	61	31.79	11.40	0.90	3.68
	White, Non-Hispanic	8427	61	37.33	9.68	0.86	3.65
	Native Hawaiian/Other Pacific Islander	29	61	32.93	11.16	0.89	3.71
	Female	4958	61	35.59	10.00	0.86	3.69
	Male	5385	61	36.27	10.48	0.88	3.65
	Limited English Proficient	201	61	20.63	7.07	0.75	3.53
	Migrant	36	61	33.00	10.56	0.88	3.72
	Plan 504	159	61	34.60	9.52	0.85	3.72
	All Students	10343	61	35.94	10.26	0.87	3.67
10	Special Education	976	61	24.82	8.99	0.85	3.46
	Title 1	66	61	28.55	7.76	0.80	3.49
	Low Income	3400	61	31.31	10.24	0.88	3.50
	American Indian or Alaskan Native	1058	61	27.00	9.91	0.88	3.48
	Asian	96	61	39.02	9.40	0.86	3.50
	Hispanic	384	61	31.54	9.77	0.87	3.50
	Black or African American	113	61	31.61	8.57	0.83	3.54
	White, Non-Hispanic	8468	61	36.19	9.78	0.87	3.49
	Native Hawaiian/Other Pacific Islander	33	61	37.55	8.60	0.84	3.48
	Female	4920	61	34.46	9.81	0.87	3.51
	Male	5232	61	35.59	10.53	0.89	3.49
	Limited English Proficient	128	61	19.20	6.02	0.68	3.41
	Migrant	21	61	30.57	9.26	0.85	3.58
	Plan 504	187	61	34.66	9.86	0.88	3.48
	All Students	10152	61	35.04	10.20	0.88	3.50

**Table N-2. 2013–14 MontCAS: Reliabilities
by Reporting Category—Science**

Grade	Item Reporting Category*	Number of Items	Raw Score			Alpha	SEM
			Maximum	Mean	Standard Deviation		
4	1	11	14	7.77	2.83	0.62	1.75
	2	14	14	10.07	2.39	0.59	1.54
	3	14	14	8.53	2.78	0.65	1.65
	4	11	14	8.29	2.55	0.60	1.61
	5	2	2	1.16	0.76	0.35	0.61
	6	3	3	1.94	0.88	0.30	0.74
8	1	14	14	8.56	2.91	0.67	1.67
	2	14	14	9.07	2.72	0.63	1.66
	3	11	14	8.03	3.00	0.57	1.97
	4	14	14	8.53	2.61	0.60	1.64
	5	2	5	1.76	1.30	0.24	1.14
10	1	14	14	9.43	2.77	0.69	1.53
	2	11	14	7.16	3.04	0.61	1.90
	3	14	14	8.13	2.76	0.64	1.64
	4	11	14	6.90	2.46	0.55	1.65
	5	3	3	2.04	0.87	0.40	0.68
	6	2	2	1.39	0.69	0.25	0.60

* Please note: 1 – Science Investigations; 2 – Physical Science; 3 – Life Science; 4 – Earth/Space Science; 5 – Impact on Society; 6 – Historical Development.

APPENDIX O—INTERRATER AGREEMENT

**Table O-1. 2013–14 MontCAS: Item Level Interrater Consistency Statistics—
Science**

<i>Grade</i>	<i>IREF</i>	<i>Number of</i>		<i>Percent</i>		<i>Correlation</i>	<i>Percent of Third Scores</i>
		<i>Score Categories</i>	<i>Responses Scored Twice</i>	<i>Exact</i>	<i>Adjacent</i>		
4	257333	5	219	65.3	29.68	0.81	5.02
	159638	5	220	67.73	28.18	0.79	4.09
8	158487	5	223	59.64	34.53	0.86	5.83
	121235	5	218	60.09	34.4	0.75	5.5
10	52993	5	192	74.48	23.96	0.92	1.56
	56042	5	197	73.1	21.83	0.75	5.08

APPENDIX P—DECISION ACCURACY AND CONSISTENCY RESULTS

Table P-1. 2013–2014 MontCAS: Summary of Decision Accuracy (and Consistency) Results by Content Area and Grade—Overall and Conditional on Performance Level

<i>Content Area</i>	<i>Grade</i>	<i>Overall</i>	<i>Kappa</i>	<i>Conditional on Level</i>			
				<i>Novice</i>	<i>Nearing Proficiency</i>	<i>Proficient</i>	<i>Advanced</i>
Science	4	0.79 (0.71)	0.55	0.77 (0.64)	0.75 (0.66)	0.81 (0.76)	0.82 (0.66)
	8	0.76 (0.67)	0.51	0.78 (0.67)	0.70 (0.61)	0.79 (0.73)	0.77 (0.60)
	10	0.74 (0.65)	0.52	0.80 (0.71)	0.72 (0.64)	0.64 (0.54)	0.86 (0.74)

**Table P-2. 2013–2014 MontCAS: Summary of Decision Accuracy (and Consistency) Results
by Content Area and Grade—Conditional on Cutpoint**

Content Area	Grade	Novice / Nearing Proficiency			Nearing Proficiency / Proficient			Proficient / Advanced		
		Accuracy (Consistency)	False		Accuracy (Consistency)	False		Accuracy (Consistency)	False	
			Positive	Negative		Positive	Negative		Positive	Negative
Science	4	0.96 (0.95)	0.02	0.02	0.90 (0.87)	0.05	0.04	0.92 (0.89)	0.05	0.02
	8	0.95 (0.93)	0.02	0.03	0.90 (0.86)	0.05	0.04	0.91 (0.88)	0.06	0.03
	10	0.93 (0.91)	0.03	0.04	0.89 (0.85)	0.06	0.05	0.91 (0.88)	0.06	0.03

APPENDIX Q—SAMPLE REPORTS



MontCAS

Criterion-Referenced Test (CRT-Science)

Student Report

2014

Letter from Superintendent

Dear Parents/Guardians:

The Montana Comprehensive Assessment System (MontCAS) Criterion-Referenced Test (CRT) is the state's measure of student performance on the state content standards which establish goals for what all students should know and be able to do.

This year, Montana students will participate in a field test or a "test of the test" of a new online assessment in English and Math in grades 3-8 and grade 11 with the exception of qualifying students with disabilities. A field test is used to evaluate the testing software, ensure the quality of test questions, and evaluate the effectiveness of the test administration and training materials. This new assessment matches new state content standards in English and Math. Because a field test is a "test of the test", students, parents, schools and the state will not receive scores.

Students in grades 4, 8, and 10 continue to be assessed using the current paper-and-pencil version of the CRT in Science. The CRT Science assessment contains multiple-choice questions and constructed response items. The constructed response items give students the opportunity to explain answers and solve problems using multiple strategies.

This report shows how your student performed on the March 2014 Science CRT. The results of this standard-based assessment are reported in four performance levels: Advanced, Proficient, Nearing Proficiency, and Novice. While some students may not yet meet the standards, keep in mind that the standards are rigorous and challenging. The staff at your school will be able to provide further information about your student's performance on the CRT.

This is only one measure of student performance and should be viewed in the context of the student's local programs and other measures. I encourage you to contact your student's school to begin a conversation that will support your student's success.

Sincerely,

Denise Juneau
Montana Superintendent of Public Instruction
Montana office of Public Instruction
PO Box 202501
Helena, Montana 59620-2501
<http://www.opi.mt.gov>

What can you do to help your student?

It is important to support your student in his or her studies now and throughout his or her future education.

Here are some tips for supporting your student in the completion of his or her schoolwork:

- Have regular discussions with your student's teacher(s) to see what you can do at home to support your student's work in school, such as making sure homework is done.
- Discuss with your student the subjects in which he or she needs improvement. Talk about whether there has been a noticeable improvement. If not, find out why.
- Ask your student to explain what he or she is studying. These conversations help you to follow your student's progress and help your student to remember what he or she has learned.
- Make sure your student gets enough rest, eats properly, and arrives at school on time every day. Send your student to school prepared to learn.

What is the MontCAS Criterion-Referenced Test (CRT-Science)?

The Montana Comprehensive Assessment System (MontCAS) was developed in accordance with the following federal laws: Title 1 of the Elementary and Secondary Education Act (ESEA) of 1994, P. L. 103-382, and the No Child Left Behind Act (NCLB) of 2001.

The CRT test questions are based on, and aligned to, Montana's content standards, benchmarks, and grade-level expectations in Science. Montana educators worked with the Montana Office of Public Instruction and Measured Progress to develop test questions that assess how well students have met Montana grade-level expectations for each content area.

MontCAS CRT scores are intended to be useful indicators of the extent to which students have mastered the materials outlined in the Montana Science content standards, benchmarks, and grade-level expectations.

Who must take the CRT-Science?

All classroom students in grades 4, 8, and 10 enrolled for 180 hours or more in an accredited public or private Montana school are required to participate.

What subjects were tested in spring 2014?

Science

Grades 4, 8, and 10

What types of questions are on the CRT-Science?

- Multiple-choice questions: Students choose the correct answer from four options and receive one point for each correct answer and zero points for an incorrect answer.
- Constructed-response questions: Students are asked to explain and/or make a chart, table, diagram, illustration, or graph to support their answer. Each answer receives zero to four points.

How are the CRT-Science results used?

MontCAS CRT-Science test results are used for the following purposes:

- to assist educators in planning improvements to curriculum and instruction
- to determine whether schools are helping their students meet the state content standards

Where can you find more information?

Where you can find more information:
<https://data.opi.mt.gov/opireportingcenter>

Montana requirements for the participation of students with disabilities on the CRT-Science:
<http://www.opi.mt.gov/Curriculum/MontCAS>

OPI contact:
Judy Snow, State Assessment Director
406-444-3656
jsnow@mt.gov

Your student's performance level and score in each content area

Display of scores and probable range of scores

In the figure below your student's performance is displayed. For each subject, the left column lists the possible performance levels with the scores needed to achieve those levels. The center column is your student's performance where the black bar is their score and the small grey bar is the range of scores they might have achieved had they taken the test multiple times. The right hand column is the percentage of students that achieved each performance level on the CRT-Science across the state.

Example:
 Your student's score → 240 ← Range of likely scores if your student took the test many times

			Science					
Performance Levels	Student	State Percentage	Performance Levels	Student	State Percentage	Performance Levels	Student	State Percentage
Advanced			Advanced			Advanced 282-300		21%
Proficient			Proficient			Proficient 250-281	262	47%
Nearing Proficiency			Nearing Proficiency			Nearing Proficiency 225-249		26%
Novice			Novice			Novice 200-224		6%

In 2014 Students were assessed only in **Science** on the CRT. Please see the letter on the front of this report for more information.

Your student's Science Scaled Score is **262** which is at the **Proficient Level**. Your student's possible range of scores is from 254 to 270.

Students at this level demonstrate a solid understanding of challenging subject matter and are able to:

- With direction, safely complete a simple investigation by asking questions using identified variables, use appropriate tools, communicate results, and identify that observation is a key inquiry process used by Montana American Indians.
- Select and use tools for simple measurement of solids, liquids, and gases, identifying properties of each state of matter and describing and modeling characteristics of and changes within basic physical and mechanical systems.
- Identify attributes of biotic (living) and abiotic (non-living) objects, including classification based on similarities and differences, basic structure and function, and processes of each system.
- Identify and accurately illustrate Earth's features, locating several observable changes of those features.
- Identify interactions among technology, science, and society.
- Discuss scientific information related to current events and local problems.
- Identify the historical significance of scientists, identify the impact of their discoveries on humans today, and identify influences of science and technology on the development of Montana American Indian cultures.
- Identify examples of Montana American Indian contributions to scientific and technological knowledge.

Scores on Montana Content Standards

CRT-Science results are reported for Montana Content Standards in Science to provide standard-specific information about the student's achievement. The results can be used to show the student's relative performance on the standards within a content area.

Science	Total Possible Points on the Test	Points Earned by Your Student	Range of Points Earned by Students Who Have Achieved Proficiency in the State
1. Scientific Investigations	14	11	3-14
2. Physical Science	14	6	4-14
3. Life Science	14	8	3-14
4. Earth/Space Science	14	11	3-14
5. Impact on Society	Subscores are not reported for this standard.		
6. Historical Development	Subscores are not reported for this standard.		

MontCAS CRT

School: Demonstration School 2
System: Demonstration District A
Grade: 04
Spring 2014

Science

School Summary Report

Confidential

I. Distribution of Scores

Perf. Level	Scores	School			System			State		
		N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.
Advanced	297–300	1	6	19	3	8	23	722	7	21
	293–296	1	6		2	5		232	2	
	290–292	0	0		0	0		300	3	
	286–289	0	0		1	3		315	3	
	282–285	1	6		3	8		671	6	
Proficient	276–281	1	6	44	3	8	51	748	7	47
	269–275	1	6		4	10		1,259	12	
	263–268	2	13		4	10		848	8	
	256–262	2	13		6	15		1,523	14	
	250–255	1	6		3	8		714	7	
Nearing Proficiency	245–249	0	0	25	1	3	13	913	8	26
	240–244	2	13		2	5		577	5	
	235–239	2	13		2	5		680	6	
	230–234	0	0		0	0		382	4	
	225–229	0	0		0	0		287	3	
Novice	220–224	0	0	13	1	3	13	313	3	6
	215–219	1	6		2	5		137	1	
	210–214	0	0		1	3		103	1	
	205–209	1	6		1	3		37	0	
	200–204	0	0		0	0		84	1	

Results are suppressed when less than ten (10) students were assessed.

II. Subtest Results

Science		Possible Points	Average Points Earned		
			School	System	State
Total Points		61	36	39	38
Standards	1. Scientific Investigations	14	7	8	8
	2. Physical Science	14	10	10	10
	3. Life Science	14	8	9	9
	4. Earth and Space Science	14	8	9	8
	5. Impact on Society	Sub scores are not reported for this standard			
	6. Historical Development	Sub scores are not reported for this standard			

CRT Performance Level Descriptors

Advanced (282–300)

This level denotes superior performance.

Proficient (250–281)

This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225–249)

This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200–224)

This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

III. Results for Subgroups of Students

Reporting Category	School					System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students	16	13	25	44	19	39	13	13	51	23	10,845	6	26	47	21
Gender															
Male	10	10	30	40	20	22	9	14	45	32	5,516	6	25	47	22
Female	6	*	*	*	*	17	18	12	59	12	5,328	6	28	47	20
Ethnicity															
American Indian or Alaskan Native	3	*	*	*	*	5	*	*	*	*	1,498	19	47	30	5
Asian	1	*	*	*	*	1	*	*	*	*	130	5	24	44	28
Hispanic	1	*	*	*	*	3	*	*	*	*	432	8	35	45	12
Black or African American	1	*	*	*	*	1	*	*	*	*	157	10	37	45	8
Native Hawaiian or Other Pacific Islander	1	*	*	*	*	1	*	*	*	*	45	9	42	31	18
White	9	*	*	*	*	28	11	11	46	32	8,582	4	22	50	24
Special Education	2	*	*	*	*	6	*	*	*	*	1,166	20	41	33	6
Students with a 504 Plan	0	*	*	*	*	1	*	*	*	*	82	10	12	57	21
Title I (optional)	0	*	*	*	*	0	*	*	*	*	19	11	63	16	11
Tested with Standard Accommodation	1	*	*	*	*	6	*	*	*	*	1,068	19	44	32	5
Tested with Non-Standard Accommodation	0	*	*	*	*	1	*	*	*	*	5	*	*	*	*
Alternate Assessment	If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report														
Migrant	0	*	*	*	*	1	*	*	*	*	29	7	21	62	10
Gifted/Talented	1	*	*	*	*	2	*	*	*	*	558	0	1	28	70
LEP/ELL	1	*	*	*	*	1	*	*	*	*	388	33	51	14	2
Former LEP Student	1	*	*	*	*	1	*	*	*	*	172	10	50	35	5
LEP Student Enrolled for First Time in a U.S. School	1	Performance levels are not reported for 1st year LEP students													
Free/Reduced Lunch	6	*	*	*	*	13	23	8	46	23	4,924	10	35	43	11

*Less than ten (10) students were assessed



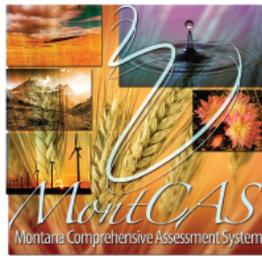
CONFIDENTIAL

Student Name
LUCIANA ALBANESE

Longitudinal Data Report

Year	Enrolled Grade	School Name	Administration	Test Name	Content Area	Score	Performance Level
0910	04	Demonstration School 1	MontCAS CRT	Grade 04 Mathematics	mat	240	Nearing Proficiency
0910	04	Demonstration School 1	MontCAS CRT	Grade 04 Reading	rea	284	Proficient
0910	04	Demonstration School 1	MontCAS CRT	Grade 04 Science	sci	257	Proficient
1011	04	Demonstration School 3	MontCAS CRT	Grade 04 Mathematics	mat	273	Proficient
1011	04	Demonstration School 3	MontCAS CRT	Grade 04 Reading	rea	285	Proficient
1011	04	Demonstration School 3	MontCAS CRT	Grade 04 Science	sci	265	Proficient
1112	04	Demonstration School 1	MontCAS CRT	Grade 04 Mathematics	mat	300	Advanced
1112	04	Demonstration School 1	MontCAS CRT	Grade 04 Reading	rea	300	Advanced
1112	04	Demonstration School 1	MontCAS CRT	Grade 04 Science	sci	275	Proficient
1213	04	Demonstration School 1	MontCAS CRT	Grade 04 Mathematics	mat	244	Nearing Proficiency
1213	04	Demonstration School 1	MontCAS CRT	Grade 04 Reading	rea	255	Proficient
1213	04	Demonstration School 1	MontCAS CRT	Grade 04 Science	sci	233	Nearing Proficiency
1314	04	Demonstration School 1	MontCAS CRT	Grade 04 Science	sci	262	Proficient

Note: This report returns as many years of data as are available for this student beginning with 06-07.



Science Item Analysis Summary

System: Demonstration District A

School: Demonstration School 1

Grade: 04

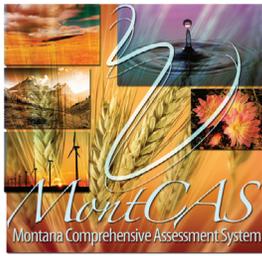
Date: 9/11/2014 8:16:22 AM

Multiple Choice

Released Item	Standard	Correct (#)	A (#)	B (#)	C (#)	D (#)	IR (#)	Correct Response
1	3	14	14	0	2	1	0	A
2	5	12	3	12	0	2	0	B
3	2	9	2	3	9	3	0	C
4	4	14	3	0	0	14	0	D
5	5	6	2	6	4	5	0	B
6	2	10	10	2	0	5	0	A
7	1	11	1	11	4	1	0	B
8	4	12	3	0	12	2	0	C
9	4	14	2	1	14	0	0	C
10	1	8	6	0	3	8	0	D
11	4	12	2	3	12	0	0	C
12	4	11	1	11	2	3	0	B
13	2	16	1	16	0	0	0	B
14	3	11	4	11	1	1	0	B
15	3	6	4	6	1	6	0	D
16	3	15	1	15	1	0	0	B
17	2	11	11	3	2	1	0	A
18	1	11	5	0	1	11	0	D
19	1	9	2	9	3	3	0	B
20	3	12	1	4	0	12	0	D
21	3	12	1	3	12	1	0	C
22	2	13	1	3	0	13	0	D
23	2	11	2	1	3	11	0	D
24	4	11	2	3	11	1	0	C
25	1	14	14	1	2	0	0	A
26	6	14	0	2	14	1	0	C

Constructed Response

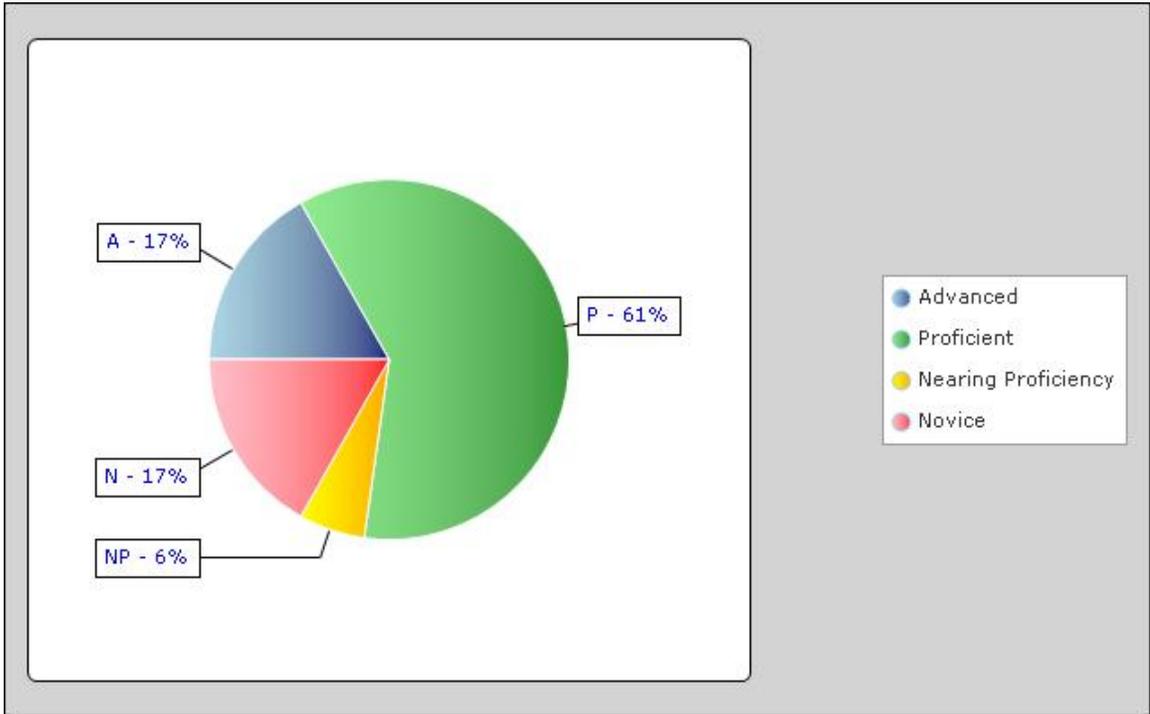
Released Item	Standard	Point Value	Average Score
27	4	4	1.1



Performance Level Summary

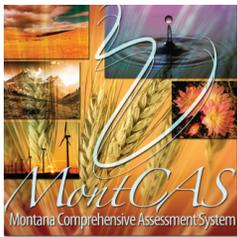
System: Demonstration District A
School: Demonstration School 1
Grade: 04
Date: 9/11/2014 8:15:56 AM

Science



Performance Level	Count	Percentage %*
Advanced	3	17
Proficient	11	61
Nearing Proficiency	1	6
Novice	3	17

*Percentages may not total exactly 100% due to applied rounding.



C o n f i d e n t i a l

Roster and Item-Level Report

Science

System:	Demonstration District A
School:	Demonstration School 1
Grade:	04
Date:	9/11/2014 8:15:25 AM

Page: 1 of 1

		Released Items																											Total Test Results							
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Points Earned by Standard on CRT				Total Points Earned on CRT	Scaled Score	Performance Level	
		Content Standard	3	5	2	4	5	2	1	4	4	1	4	4	2	3	3	3	2	1	1	3	3	2	2	4	1	6	4	Standard 1	Standard 2	Standard 3				Standard 4
		Depth of Knowledge Code	2	2	2	2	2	3	2	2	1	3	1	2	1	2	2	1	1	2	1	2	2	2	2	1	2	2	2							
Item Type	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	CR									
Correct MC Response	A	B	C	D	B	A	B	C	C	D	C	B	B	B	D	B	A	D	B	D	C	D	D	C	A	C										
Name/Student ID	Total Possible Points	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	14	14	14	14	61			
ALBANESE, LUCIANA	D04100015	C	+	A	+	D	D	+	+	+	+	+	+	+	B	+	D	+	+	+	+	+	C	+	C	+	1	11	6	8	11	39	262	P		
CACERESUMANA, BRYAN	D04100030	C	+	D	+	+	D	+	+	+	+	B	+	+	+	+	+	+	+	+	+	+	A	+	+	+	2	10	11	12	11	48	285	A		
CAMPBELL, ANDREA	D04100038	D	+	B	+	C	+	C	+	+	+	D	+	A	A	+	+	+	A	+	+	+	+	+	+	1	8	12	8	10	40	264	P			
CLOUGHERTY, FINBAR	D04100009	+	+	B	+	+	+	+	+	+	+	D	+	A	B	+	+	+	D	+	B	A	+	+	+	3	10	11	10	11	47	282	A			
COFFARO, SARA	D04100039	+	+	+	A	D	D	+	A	+	A	B	C	A	+	B	+	+	+	C	+	+	+	+	+	0	7	9	6	5	31	246	NP			
COMBS, DANIELLE	D04100044	+	A	+	+	C	B	+	+	+	A	+	+	+	A	+	+	A	D	+	+	B	+	B	+	0	6	10	9	8	36	256	P			
FERGUSON, MALIK	D04100035	+	+	+	+	D	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	B	+	0	11	13	14	9	52	298	A			
FINLAY, ASHLEE	D04100028	+	+	D	+	+	+	D	D	+	+	+	+	+	+	+	+	+	+	D	+	+	B	+	+	2	7	7	11	11	40	264	P			
HAIGH, MARSHA	D04100007	+	A	D	+	D	+	C	+	B	A	+	+	+	A	+	+	+	+	B	B	+	+	+	1	6	9	9	9	35	254	P				
HAMMOND, JEDEDIAH	D04100024	+	+	+	+	C	+	+	D	+	A	A	+	+	D	+	A	B	+	A	A	+	+	+	0	7	13	7	7	37	258	P				
LECLERC, ALEXANDR	D04100046	+	+	+	+	+	D	+	+	+	+	+	+	+	B	+	C	A	D	+	B	+	C	+	2	7	9	9	12	42	269	P				
LEONARD, ZACHARY	D04100042	+	+	+	+	D	+	+	+	+	C	+	C	+	+	+	+	+	C	+	B	+	+	A	B	2	5	12	12	9	42	269	P			
LUNN, SUSAN	D04100040	+	+	+	+	+	B	A	+	A	+	+	+	+	+	B	+	B	+	+	B	+	+	+	2	9	11	10	11	46	279	P				
MAJORS, DALTON	D04100041	+	A	A	A	C	+	C	A	A	C	A	A	+	A	+	+	B	A	C	B	A	B	A	0	4	4	5	3	18	218	N				
ROBINSON, SARAH	D04100027	+	D	B	A	A	+	C	A	+	A	B	+	+	A	A	C	C	A	+	B	D	A	B	0	4	4	4	3	16	213	N				
ROJAS, MICHAEL	D04100003	+	D	+	+	D	+	+	+	+	A	+	D	+	+	A	+	+	C	+	+	+	+	+	2	7	13	11	10	43	271	P				
RUDDY, JASMINE	D04100047	C	+	+	A	C	D	C	+	D	+	+	C	D	D	D	+	C	A		B	A	A	A	0	3	5		4	20	223	N				
SANTOS, JESSABEL	D04100004	+	D	+	+	+	+	+	+	+	B	A	+	+	+	+	+	B	+	+	+	B	+	+	D	1	11	12	10	9	44	274	P			
SCOTT, KENDALL	D04100026	+	+	+	+	+	+	+	+	+	+	+	+	+	+	B	+	+	+	+	+	+	+	+	2	14	13	11	12	54	300	A				
SELLERS, NOAH	D04100031																								B	0	0	0	0	0	200	DNP				
SMITH, ALEX	D04100025	+	+	B	+	A	+	+	+	+	C	+	D	+	C	C	+	+	A	+	+	+	+	2	7	9	11	9	39	262	P					
Released Item Number		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27								
Percent Correct/Avg. Score: Group		82	71	53	82	35	59	65	71	82	47	71	65	94	65	35	88	65	65	53	71	71	76	65	65	82	82	1.1	7.6	9.7	9.2	8.8				
Percent Correct/Avg. Score: School		82	71	53	82	35	59	65	71	82	47	71	65	94	65	35	88	65	65	53	71	71	76	65	65	82	82	1.1	7.6	9.7	9.2	8.8				
Percent Correct/Avg. Score: System		65	70	68	78	35	68	65	76	73	46	65	62	97	62	38	86	62	73	46	68	59	76	65	70	76	76	1.0	7.7	10.1	8.8	8.5				
Percent Correct/Avg. Score: State		62	67	64	85	50	73	71	73	71	50	64	51	95	58	47	82	68	68	42	63	48	58	56	53	57	68	1.1	7.8	10.1	8.6	8.3				