Maryland's Accountability Assessment Program

A Brief History and Generic Overview of the Test Development Process

December 2008

Table of Contents

Tables

Appendices

Executive Summary

This paper provides a brief history of Maryland's accountability assessment program, an overview of the essential components which over the years have substantively contributed to the success of these assessments, and an overview of the typical procedures required for the design and development of high quality tests that satisfy national test development standards. The assessment programs summarized include commercially developed standardized tests, the Maryland Functional Testing Program (MFTP), the Maryland School Performance Assessment Program (MSPAP), the High School Assessments (HSA), and the Maryland School Assessments (MSA).

The essential components include curriculum, requests for proposals (RFPs), participatory relationships with vendors, educational stakeholder involvement at every level, oversight of the process and results, transparency, and continuous improvement.

The basic process of test development involves defining the purposes of the tests, planning, consulting with multiple contractors, program management, test item development and review, publishing of test materials, administration, security monitoring, scoring, standard setting, reporting of results, research, policy development, and communications with all stakeholders. How assessments are adapted to meet the needs of students with disabilities or limited English proficiency is also addressed, as well as processes to ensure the tests are accessible to an increasingly diverse student population.

The increase in student performance on these assessments over the years is indicative of the improved instruction seen across the State, and a changing assessment system that has been adapted to become more demanding as teachers and students have risen to meet each new level of challenge. This is particularly the case with traditionally underperforming subgroups of students who have benefitted from improved aligned curriculum delivered by increasingly better qualified and better educated teachers.

Brief History of Assessment in Maryland

Standardized Tests

Standardized tests have been used for many years in education, generally as an efficient way for states to provide information to schools and parents on student achievement. These tests are often norm-referenced assessments that compare student performance to a national norm group. They consist of multiple choice items and are still used in some states. Some of the limitations of these national tests are that they are not aligned to any state's curriculum, the norms are applied for years, ultimately resulting in large numbers of students scoring "above the norm," and since the same items are administered year after year, teachers become very familiar with the test. An increased desire for accountability and the standards movement drove educators to look for assessments that were aligned to their curriculum, and allowed for criterion-referenced interpretation of scores. In Maryland, the first battery of such tests was the Maryland Functional Testing Program.

Maryland Functional Testing Program (MFTP)

The purpose of this testing program was to assess students' basic competencies in reading mathematics, writing and citizenship. Passing the tests was required for high school graduation. The reading, mathematics, and citizenship tests were composed of multiple-choice items; the

writing test consisted of two prompts. At the beginning of the program, students were tested for the first time in Grade 9, and continued to take the test until they passed. By the end of the program, students were taking and passing the tests in grade 6.

Maryland School Performance Assessment Program (MSPAP)

The Maryland School Performance Assessment Program was an award winning performance assessment administered to students in grades 3, 5, 8. It was composed of theme based tasks that consisted of only constructed response items that integrated the measurement of reading, language usage, writing, mathematics, science, and social studies. MSPAP was the first and only State test to use a "matrix-sampled design." This design was a measurement format in which a large set of tasks were organized into a number of relatively smaller sets or forms which were randomly assigned to test takers so that no student responded to all of the items. Maryland's design consisted of three similar sets or forms, each containing one-third of the total test. Because a given student answered only one-third of the questions, that student responded to too few items to receive a stable individual score. Therefore, the scores from the three forms were aggregated for schools and districts so that they would be stable for program evaluation and school improvement.

MSPAP also had other unique features. Students took the test for approximately two-and-a-half hours per day for five week days. Students had the option to respond to tasks by drawing, graphing, or writing. Students also worked in small groups to conduct short experiments and used the data collected as the basis for their answers. The assessment modeled good instruction by engaging students and by including "real world" scenarios to assess students' ability to apply higher order thinking skills to actual problems.

Maryland School Assessments (MSA)

In 2002, the Maryland State Department of Education (MSDE), in order to conform to the requirements of the new Federal program "No Child Left Behind," retired its award-winning Maryland School Performance Assessment Program and began the development of a new testing program known as the Maryland School Assessments (MSA). The new program, like its predecessor, is based on the Voluntary State Curriculum, which sets reasonable grade level academic standards for what teachers are expected to teach and for what students are expected to learn in public schools. Maryland has administered the reading and mathematics tests annually to students in grades 3, 5 and 8 since 2003 and to students in grades 4, 6, 7 since 2004. Maryland has administered a science test annually in grades 5 and 8 only since 2007.

In order to quickly and efficiently complete the transition to NCLB testing requirements, most notably the need for individual student scores, the MSA was developed using an item-sharing concept. Existing items from a national norm-referenced test were evaluated and aligned to Maryland's standards. Those that aligned to the standard and met the appropriate psychometric properties became part of the new MSA. The state standards not assessed by the norm-referenced test items were assessed with new items written using Maryland's normal item-writing process, and included constructed response items. Thus the MSA scores were generated from a mix of norm-referenced test items augmented with items written by Maryland educators, all of which assess Maryland's content standards.

By 2007, it became clear that the use of the norm-referenced items was causing confusion in classrooms and unnecessary extended testing time. Because the norm-reference items could only be administered as part of the entire norm-referenced test, students were taking many items that in fact did not count toward their MSA score and were not aligned to Maryland

standards. In 2008, these items were removed and replaced by items developed by Maryland educators to measure the same standards in the same way, aligned to Maryland standards.

Now each student receives an MSA score that categorizes his/her performance according to one of three achievement standards: Basic, Proficient, or Advanced (see Performance Level Standards). The MSAs are peer reviewed by the US Department of Education and must satisfy the rigorous requirements of the federal NCLB Act.

High School Assessments (HSA)

In 1992 the Maryland State Board of Education received the recommendations of a State task force for end-of-course assessments for a set of core high school courses.

By 1994, MSDE engaged ETS to conduct public forums around the State on the proposal for High School Assessments (HSAs), with an implementation plan prepared by the following year. In 1995, the State Board received a plan to phase out the functional test graduation requirement and require students to pass the HSA tests. Subsequently, in 1996, the Core Learning Goals (CLGs) were released to the local school systems for incorporation into their own local curriculum and the State Board authorized the development of the exams so they could become a requirement for graduation. The tests were first administered in 2000 and standards were set to define passing. In 2009 school systems will have been working with these standards for 13 years, of the entire time the graduating class has been in school.

Like the MSAs, the HSAs are high-stakes, standards-based tests designed to measure the Voluntary State Curriculum known as the Core Learning Goals (CLGs). Unlike the MSAs, the HSAs are end-of-course tests which are administered to students upon completion of courses in English2, government, algebra/data analysis, and biology. The HSAs are composed of both selected response (SR) items and constructed response (CRs) items. The algebra/data analysis test also contains student produced response (SPR) items. Unlike the MSPAP program, each test yields scores for individual students, in addition to scores for the state, school systems, and schools. The tests in English2, government, algebra/data analysis, and biology will be graduation requirements for all students in Maryland public schools beginning with the class of 2009. The tests in English2, algebra/data analysis, and biology are also used to meet the No Child Left Behind (NCLB) testing requirement and are peer reviewed by the US Department of Education and must satisfy the rigorous requirements of the federal NCLB Act.

Essential Components

Across all of these testing programs, there have been some key components that have provided stability, integrity, quality, and acceptance for Maryland's tests. The test development process has been well-tested over time, and has allowed early approval from the United States Department of Education (USDE) for Maryland's NCLB assessment system. While Maryland has used the same basic process over time, modifications have been instituted to accommodate new testing requirements and purposes, as well as to take advantage of new methodology and formats. The following have been identified by the Government Accountability Office as the components that have contributed significantly to the quality and reputation of Maryland's assessment programs.

Curriculum

The core of a quality assessment program is that it is aligned with a well articulated and defined curriculum. Maryland has developed curricula in partnership with many Maryland teachers at each grade level and in each content area. The process also includes activities to benchmark Maryland's curriculum to national content standards. The curriculum is reviewed by outside groups (such as Achieve) and is only adopted after State Board approval. To allow time for school systems to transition their own curricula and adjust their instructional programs, a new curriculum is implemented years prior to the testing programs with which it is aligned. The current curriculum, the Voluntary State Curriculum (VSC) is supported by a multitude of resources for school and teachers. Many of these are available on the school improvement website: www.mdk12.org.

Details on why the curriculum is a critical element in the test development process, and the benchmarking to national standards are presented later in this paper in the steps for test development.

Requests for Proposals

The complex process of test development requires a well organized approach and significant expert involvement. Therefore, Maryland partners with national testing companies to accomplish these tasks. Successful processes and products begin with original requests for proposals (RFPs) that are written by MSDE with specific requirements, expectations and deliverables. Multiple contractors seeking to work with Maryland submit proposals in response to these requests. Following a critical evaluation of corporate competence and capacity, the contract is awarded to the most capable vendor.

A given contractor may work with Maryland for several years because it takes a minimum of three years to develop and implement an assessment. At any given time, staff is working on a different phase of tests for three different years of administration. For example, in year one new items are being written for possible inclusion on the assessment in year three, field tested items are being evaluated and forms are being constructed for the assessment in year two, and the current years' test is being printed and delivered to schools for administration, scoring and reporting.

Participatory Relationship with Vendors

While some states delegate their development, administration and scoring process to their contractor and await results, Maryland has an on-going and constant partnership with its testing contractors. Working side by side on every aspect—MSDE is aware of every challenge, and is a part of every solution. The goal is to ensure no surprises, therefore contractor decisions are vetted through MSDE staff and sometimes other experts. Staff and contractors participate in weekly (and often more) conference calls and have day-long meetings to plan implementation. The result is a truly Maryland assessment program that meets the specific needs of Maryland's public schools.

Educational Stakeholder Involvement at Every Level

Maryland teachers and educators are involved in the test development process at every step from curriculum development through item writing, item bias reviews, developing scoring materials, to setting performance standards. The educators working on the test development process are inclusive of all groups: advocates for English language learners, students with disabilities, racial/ethnic groups, and represent the entire state (all LEAs). Hundreds of

Maryland teachers have participated in the development of Maryland's assessment programs. More details are included in the discussion of the test development process.

Oversight of (Outside Validation of) Process and Results

National Psychometric Council

Maryland was one of the first states to have a National Psychometric Council (NPC)—a group of independent professionals, expert in statistics and measurement who serve as advisors to contractors and MSDE and provide final outside approval of statistics and results for assessment design and results. Maryland assessments used for NCLB accountability are also rigorously reviewed by the federal peer review process.

U.S. Department of Education Peer Review Process

To determine whether States have met NCLB standards and assessments requirements, the U.S. Department of Education (USED) uses a peer review process involving experts in the fields of educational standards and assessments. Peers examine evidence compiled and submitted by each State that is intended to show that its assessment system meets NCLB requirements. Such evidence may include, but is not limited to, reports of alignment and validation studies; written policies for providing accommodations for students with disabilities and limited English proficient students (LEP); written policies on native-language testing of LEP students; score reports showing disaggregation of student achievement data by the statutorily specified student subgroups, State statutes, State regulations, test administration manuals, board resolutions, and technical assessment reports.

The state's evidence is sent by USED to each peer reviewer weeks in advance of a review meeting to allow each peer to thoroughly and independently review it before they meet on site in three to four member peer review teams. Peer reviewers use a *Guidance* document composed of detailed questions as a framework (2007) to record their impressions of the degree to which the State's final assessment system complies with the requirements of Title I. Their collective evaluation of the assessment system serves two purposes: a technical assistance tool to support improvements in the system; and recommendations that inform the decision of the Assistant Secretary for Elementary and Secondary Education regarding approval of the State's assessment system.

Role of Peer Reviewers and Review Teams

The on-site peer review team prepares a detailed report based on its examination of the materials submitted by the State. In each team, one peer serves as the designated team leader; who is responsible for composing peer notes that are clear, complete, and delivered to USED staff at the end of the review meeting. The peer reviewers are responsible for providing feedback to each State that is informative and is consistent with professional standards and best practice. Generally, if changes in a State assessment system are required in order to meet Title I requirements, peer reviewers present options rather than prescriptive instructions.

A USED staff person, assigned as a resource to each team, is responsible for assisting the review team in obtaining adequate and appropriate information from the State prior to the review meeting; contacting the State during the review meeting to obtain clarification or additional information needed by the reviewers; securing resources needed to support the team during the meeting; and accurately reporting the review team's deliberations as USED determines the State's compliance status. USED staff may question, or even challenge, the

peer reviewers in order to promote clarity and consistency with the *Guidance*; they will not, however, impose their views or require substantive changes in the peer reviewers' report.

Transparency or Public Disclosure

Numerous resources make the assessment program transparent for diverse educational stakeholders. Documents are written by MSDE and its contractors for measurement experts, parents, students, teachers, and administrators and must be complete, current, accurate and clear. Documents typically specify the nature of the assessments, their intended use, content and skills that are measured, the procedures used for the development, administration, scoring, reporting, and interpreting the results. Some are secure and may be viewed by only those with a need to know the contents. Examples of secure documents include test specifications, test items, and the actual operational test forms.

Most documents, however, are in the public domain and can be accessed by anyone. Examples of non secure documents include the core learning goals, the voluntary state curriculum, handbooks for parents and students, test administration manuals, scoring guides, score interpretation manuals, state and school district score reports, annual disclosed test forms, technical manuals, research studies, sample items of the type included on the assessments, and the characteristics of educators who participated in item reviews, standard setting, and alignment studies.

Educators may also learn about the test development process through participation in professional development opportunities e.g., item writing, scoring of constructed response items, standard setting, and data driven decision making workshops The MSDE websites also include videos, state events calendars, fact sheets, frequently asked questions (FAQs) and answers, score reports, and other special reports

Research

The Department has a strong relationship with the Maryland Assessment Research Center for Education Success (MARCES). MARCES, a project of the Department of Measurement, Statistics, and Evaluation in the College of Education at the University of Maryland, provides support to the range of assessment activities in the State, the region and the nation by conducting basic and applied research to enhance the quality of assessment practice and knowledge. Studies and topics researched are determined each year according to MSDE needs and the federal peer review requirements.

Continuous Improvement

Maryland has a long history of willingness to listen to LEAs, to embrace new opportunities and, technologies to improve assessments to better show case the achievement of Maryland students. Recent examples include: the creation of the alternate assessments with modified achievement standards for high school students (Mod-HSAs), the elimination of items that permitted a norm-referenced interpretation from the Maryland School Assessments (MSA) to improve alignment to the curriculum and shorten testing time, the elimination of constructed response items from High School Assessments to decrease scoring costs and the turn around time from administration to reporting results to schools and students, and the use of computer assisted administrations for the Mod-HSAs and science MSAs.

Test Development

Test design and development is a complex and exacting process. Producing a test involves multiple interrelated decisions, checks and balances, and painstaking attention to detail. Each item, each form, and each score reported for every assessment undergoes intense scrutiny. Quality at every step of the process is essential and must be monitored and documented, because it builds evidence of reliability, validity, and fairness—the three most important qualities of test scores.

Reliability is the empirical degree to which test scores for a group of examinees are consistent or stable. In general, the higher the reliability, the lower the random measurement error. High reliability is an essential condition for high validity.

Validity, the most important of the three qualities, is the degree to which accumulated evidence and theory support specific interpretations of the test scores proposed by the test developers. Validity evidence can be extensive and diverse. It may be based on the test's content, test taker response processes, the test's internal structure, the test's relationship to external variables, and the intended and unintended consequences of testing.

Fairness, in the measurement context, is typically defined by four criteria: 1) absence of bias, 2) equitable treatment of test takers before, during and after testing, 3) equity in opportunity to learn the knowledge and skills measured by the test, and 4) tests composed of items which test takers of comparable ability should be able to answer correctly.

National Standards for Test Development

Maryland strictly adheres to the national criteria for best practices in test development. These criteria are defined by national standards such as the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education Standards for Educational and Psychological Tests (AERA/APA/NCME, 1999), the federal No Child Left Behind legislation (2001), the Code of Fair Testing Practices in Education (2004), the Code of Professional Responsibilities in Educational Measurement (1995), and the principals of universal design (Johnstone, 2003).

Best practice mandates that all tests which are used for important high-stakes decisions (e.g., federal accountability and state graduation requirements), such as the Maryland School Assessments (MSAs) and the High School Assessments (HSAs), must meet the most rigorous criteria in these standards. Maryland verifies that those criteria have been met by employing several ongoing procedures. These include having independent contractors conduct impartial research studies such as the alignment analysis for each test, having all assessments that are used for accountability under NCLB comprehensively and rigorously reviewed by a federal peer review process (summarized in this paper), having a technical advisory committee of measurement experts, known as the National Psychometric Council, review all test development decisions, and by seeking the ongoing feedback from diverse groups of educational stakeholders.

Definition of Test Purpose(s)

The process of developing educational tests commonly begins with a statement of the purpose(s) of the test and the scope of the content domain to be measured. All of the Maryland School Assessments—the MSAs, the Alt-MSAs, and the High School Assessments—the HSAs and the Mod-HSAs are high stakes tests that are used to track academic progress and makes accountability decisions.

Academic Content Standards

Voluntary State Curriculum

The Voluntary State Curriculum (VSC) is the document that identifies the Maryland Content Standards and aligns them with the Maryland Assessment Program. As shown in Table 1: Sample Goal, Expectation, and Assessments Limits for Grade 6 Mathematics, the curriculum documents are formatted so that each begins with content standards or broad, measurable statements about what students should know and be able to do. The next level of specificity, indicator statements, narrows the focus for teachers. Finally, the objectives provide teachers with very clear information about what specific learning should occur. When objectives are measured on the MSA in Mathematics and Reading, each is also followed by an assessment limit. Additional information on assessments limits appears later in this paper.

Table 1: Sample Goal, Expectation, and Assessment Limits for Grade 6 Mathematics

STANDARD 1.0 Knowledge of Algebra, Patterns, and Functions

TOPIC

1. PATTERNS AND FUNCTIONS

INDICATOR

1. Identify, describe, extend and create numeric patterns and functions

OBJECTIVES

1. Identify and describe sequences represented by a physical model or in a function table
2. Interpret and write a rule for a one-operation (+,-,x,÷) function

   Assessment Limits:

   - Use whole numbers or decimals with no more than two decimal places (0-10,000)
3. Complete a function table with a given two-operation rule

   Assessment Limits:

   - Use the operations of (+,-,x) numbers no more than 10 in the rule, and whole numbers (0–50)

Core Learning Goals

Core Learning Goals (CLGs) are end-of-course grade-level goals, expectations and indicators that define what a student should know and be able to do in English, Government, Algebra/Data Analysis, Biology, Geometry, Physics, Chemistry, Earth/Space Science and Environmental Science. The goal statements in English, Government, Algebra/Data Analysis, and Biology also contain assessment limits because these content areas are measured on state tests and are the four high school assessments students must pass to graduate beginning in 2009.

Assessment Limits

Assessment limits are the objectives assessed on the MSAs and HSAs at each grade level. Embedded in the Voluntary State Curriculum and Core Learning Goals, they identify the

specific skills and content that students are expected to have learned for each assessed objective. Not all objectives in the VSC have assessment limits at a given grade-level. Nonetheless, these objectives must be included in instruction because they introduce important concepts in preparation for assessed skills and content at subsequent grades.

Benchmarking to National Content Standards

The Maryland content standards assessed on the MSAs and HSAs were informed by several documents that identify the understanding, knowledge, and skills that students should acquire from preK through grade 12 produced by several national groups of content experts. The Math Content Standards were informed by the Principles and Standards for School Mathematics of the National Council of Teachers of Mathematics (1996). The Science Content Standards were informed by the National Science Education Standards of the National Research Council implemented throughout the country by the National Science Teachers Association (1996). The English Content Standards were informed by the Standards for the English Language Arts of the International Reading Association and the National Council of Teachers of English (1996). And the Government Content Standards were informed by the National Standards for Civics and Government (1994–2007).

Test Item Development

Development of test items requires an in depth knowledge of item formats, item specifications, and detailed plans and schedules for item writing, editing and review.

Item Formats and Specifications

The Maryland assessments are composed of several highly rated item formats commonly used on well known commercially developed national tests such as the SAT, the ACT, the Graduate Record Examination (GRE), the College Board Advanced Placement (AP) Examinations, and the National Assessment of Educational Progress (NAEP). These formats include the traditional Selected Response (SR), also known as Multiple-Choice (MC); the Constructed Response (CRs), which includes the Brief Constructed Response (BCR), the Extended Constructed Response (ECR); and the innovative Student-Produced Response (SPR).

Selected Response Items

Selected Response items pose a question followed by 3 or 4 suggested answers only one of which is correct or clearly best. Although some believe that SRs can only measure the simple recognition of knowledge, high quality SRs can and frequently do measure a wide range of higher order thinking skills. A major advantage of SRs is that numerous items may be used to broadly measure the content in a limited amount of testing time because the student can quickly read the question and select an answer often in less than one minute. This feature increases the test's reliability (stability of measurement) and, in turn, reduces measurement error. SRs may be "independent" (not linked to other SRs), linked alone or in clusters to a stimulus set (e.g., technical graphs or figures for lab sets), or linked alone or in clusters to a reading passage. Students can also efficiently record their answers to SRs on a machine-scorable grid. SRs are typically scored right (1 point) or wrong (0 points).

Constructed Response Items

By contrast, constructed response items, such as the brief constructed response (BCR) and the extended constructed response (ECR), ask a question to which students must compose a short or extended answer, respectively. Writing the response may require from

5 to 30 minutes per item and reduces the number of questions that might be asked in a given amount of testing time. Like the SRs, they, too, may be stand alone or linked to stimulus material.

The CRs used on the math assessments are unique. Both the BCR and the ECR have two steps: Steps A and B. Step A contributes to the content score and Step B contributes to the process score. All BCR and ECR Step A items receive a 0–1 score point range; all BCR Step B items receive a 0–2 score point range, and all ECR Step B items receive a 0–3 score point range. By contrast, on the reading assessments, the BCRs do not have steps and are scored with a general rubric with maximum values between 0–3. CRs are excellent for measuring outcomes that require students to create their answers. However, they are labor intensive and expensive to score, and increase the turn around time for the return of student scores.

Because of these limitations, and because the HSA is a high stakes graduation requirement assessment that students can re-take until they pass it, beginning in May of 2009, CRs are being phased out of the HSAs. The CRs will be replaced by SRs assessing the same content and the same level that were field tested on the 2008 test. This will allow faster turn-around time for scores so that students not passing the test can access intervention and retest in a timely fashion. Constructed response items will still remain on the MSA assessments.

Student-Produced Response

Unlike the other item formats, Student Produced Response (SPR) items are used only on the math assessments. SPRs require students to calculate answers which they have the option to record as whole numbers, fractions, or decimals on a machine-scorable grid. Because there are no supplied response choices from which students may select an answer, it is virtually impossible for students to guess the answer. Like SRs, SPRs are scored right (1 point) or wrong (0 points).

Test Development Planning Meeting

Held in September for three days, the purpose of this meeting is to review the previous year's test development process as it pertains to each content area and to plan for the upcoming year's process.

The contractor in collaboration with MSDE content staff, assigns the number of items to be developed for specific goals, expectations, and indicators. The contractor also identifies sample items for various item specifications. Best practice mandates that test items must be written to the respective goal, expectation, or indicator, and must be within the assessment limits as required by test blueprints. Items must also be written to cover a range of difficulty and subject matter. Because the number of testable goals, expectations, and indicators in the VSC is large and testing time is limited, the expectations and indicators must be sampled across time. This sampling plan for development is updated annually by the contractor and approved by MSDE prior to the start of item development for each test edition.

MSDE requires a 50% overage of items to be developed per test. For example, if 100 new field test items are needed for each content and grade, a minimum of 150 must be written per content and grade. MSDE also requires that a minimum of 100% overage of passages and stimuli needed for field test items be identified each year. Therefore, if 100 sets of

stimulus materials are needed for tests in reading, English, government and biology, a minimum of 200 sets of stimulus materials per test must be selected. The vast number of extra items and stimuli is necessary to account for attrition during the committee reviews. In addition, there is the possibility that permissions to use any copyrighted stimulus material may not be granted or available in a timely manner.

Item Writer Training

Held in October/November, the purpose of this two day meeting is for MSDE content and scoring staff to train the contractor's item writers on the VSC, CLGs, item specifications, test blueprints, and style guides and to distribute item writing assignments. All item writers who write Maryland items, the editors who review Maryland items, and the contractors' content staff assigned to work directly with Maryland content staff must attend. Historically, Maryland teachers, hired by the test development contractor, have written most test items for Maryland assessments. They know Maryland students, have experience teaching the VSC, and they are selected for their content expertise. They also represent all local school systems and all student subpopulations in the State. Once annually an additional item writer training is held on Saturdays in Maryland for Maryland-based item writers.

New Item Development Reviews

Held in December/January, the purpose of this three day meeting is to review the initial 1/3 of the test items that are newly developed. This provides the Contractor and item writers with feedback early in the process to ensure that the items are being developed to meet the Maryland criteria.

When item writers complete their assignments they submit their newly developed items to the test development contractor who carefully edits them to catch any grammatical, spelling, or punctuation flaws. They are then forwarded to MSDE in small batches every few months for review. After this review, MSDE returns the items to the contractor who edits them again according to MSDE directions.

During the editing process the contractor verifies that the item's content is aligned with the expectation, indicator and assessment limit it is intended to measure. Alignment is essential for the validity of test score interpretations. Following the alignment review, the contractor reviews the items for content accuracy, grade level appropriateness, readability, organization and comprehension, and style usage and consistency. Then, the contractor identifies all text, passages, and/or graphics that need acknowledgments and obtains print and electronic copyright and/or trademark permissions.

If the items are CRs, and the scoring contractor and the test development contractor differ, they are also reviewed and edited by the scoring contractor. Items previously field-tested are revised and re-field-tested only if the item statistics do not meet acceptable limits. Typical statistics reviewed for each item such as item difficulty, item discrimination, differential item functioning (DIF) and distractor analyses are discussed later in this paper.

Pre-Content Side-By-Side Reviews

This review is held in March at MSDE Headquarters and typically lasts 5 days. The purpose of this meeting is for MSDE and Contractor staff to review and make final edits to the items that will be going to the summer content and bias/sensitivity review meetings.

Content, Passage, and Bias/Sensitivity Reviews

About 150 educators from around the state meet in June or July in a conference center or hotel in the greater Baltimore metropolitan region to review items and item stimulus materials. One content committee meets for two days to review test items that will be field tested in the subsequent year, with a focus on grade-level content and grade-level language appropriateness. Another content committee meets for one day to review passages and stimuli that will be used for item development. A bias or sensitivity committee meets for three days to review the field tested items and the passages and stimuli to be used in item development in order to provide comments on whether the items, passages and stimuli present any topics, words or contexts that would hinder or enhance the performance of a particular group of students.

During these reviews, items are projected onto a screen using a computer which allows everyone to see the same version of the item and any changes as they are being made. This allows both the contractor and MSDE content staff to have an electronic version of the agreed upon edits at the conclusion of the meeting.

Post-Content Side-by-Side Reviews

At the conclusion of the Content, Passage, and Bias/Sensitivity Reviews, MSDE content and scoring staff and the contractor content staff meet for five days in side-by side meetings to reach agreement on the final item edits. These final edits are also conducted using projected items and electronically collected edits, with a hard copy backup.

Test Construction

Test Specifications or Blueprints

Analogous to an architectural plan for the construction of a building, precise test specifications or test blueprints are created for each content area and grade and serve as the detailed plan for the design and development of the respective assessment. The test blueprint typically mandates the test standards that should be measured, the number and/or proportion of items by item format that should measure each content and skill area, the level of cognitive demand that each item should measure, and the range of desired measurement statistics (e.g., item difficulties and item discrimination indices) of the items and the test. The blueprint may also indicate the number and sequence position of any anchor items that are used for equating, the number and sequence position of any embedded field test items, the number of items that might have associated stimulus material, and the number of items, if any, that may be linked to a single stimulus.

Alternate Test Forms

The test specifications are used to create multiple alternate test forms, also called “parallel” or “equivalent” forms, for each assessment in each content area and grade. Alternate forms are interchangeable forms that are carefully and collaboratively designed by content area and measurement experts. Each measures the same content standards using identical item formats with similar statistical properties in the same proportion. Each is administered under the same standardized conditions.

The contractor’s content and psychometric experts initially select the operational items which contribute to the student’s score. Later field test items, which do not contribute to the student’s score, will be embedded. At this stage the contractor also produces any artwork that

accompanies the items and uses special software to produce mathematical graphs, equations, inequalities and other relevant symbols. The contractor then edits manuscripts to ensure accuracy of page numbers, item numbers and item type. The test forms are then pre-equated (adjusted for difficulty across forms) prior to submission to MSDE.

As can be seen in Table 2, alternate forms permit a relatively large number of new field test items to be embedded among the operational tests for each grade. For example, for the Math – MSA there are two operational forms: Form 1 and Form 2. Each operational form has five alternate forms (Forms A through E and Forms F through K) which have identical sets of operational items. However, each of the 10 operational forms has a different set of field test items. Therefore, if 10 field test items per set are embedded on each form, 100 new field test items can be administered annually for each grade. The process permits the collection of multiple statistics that are used to evaluate the field test items without negatively impacting students because these test items do not count towards a student's score.

Table 2: Sample MSA-Math Test Structure: Grades 3 through 8

Alt. Form	Operational Item Sets		Field Test Item Sets									
	Form 1	Form 2	A	B	C	D	E	F	G	H	J	K
A	▲		●									
B	▲			●								
C	▲				●							
D	▲					●						
E	▲						●					
F		▼						●				
G		▼							●			
H		▼								●		
J		▼									●	
K		▼										●

In addition to providing multiple opportunities to field test items in operational tests, alternate forms have several other benefits. Alternate forms allow unique forms (large-print, Braille) to be adapted for special needs students, enhance test security by limiting item exposure per form, and provide students who initially fail a test required for graduation with multiple opportunities, as needed, to retake a comparable form.

Delivery of Tests to Schools

The assembly of test forms and supporting test administration manuals is required at least several months before the scheduled dates for testing to give time for printing and shipping of materials to schools. A typical timeline requires that documents be assembled in the winter months (e.g. November–December–January) in order to get them produced, printed, and packaged (February–March) for delivery to the schools in time for a spring administration (April–May). The packages of test forms are spiraled (ordered A, B, C….K) to ensure that they are randomly distributed among the students.

Test Administration

Maryland takes great care to ensure that the all assessments are administered in a fair, equitable, and standardized manner. The goal of this detailed process is to ensure that all students take the test under a uniform set of conditions so that the test results are trustworthy and can be used with confidence.

Standardized Test Administration

To promote the standardized test administration contractors develop test related examiners manuals (EM) for each content area in partnership with MSDE. EMs contain the guidelines that are used by teachers for planning and managing a standardized administration of the assessments.

In addition to the EMs, one Test Administration and Coordination Manual (TACM) is developed for use by the Local Accountability Coordinators (LAC) and building-level School Test Coordinators (STC). Included in this manual are test administration schedules, types of accommodations, procedures for administration monitoring and reporting test irregularities, as well as instructions of all permitted and prohibited activities prior to, during, and after the administration of both the computer-assisted or paper-and-pencil tests. The TACM is distributed and reviewed during test administration workshops for STCs and LACs with duplicates sent to each school with its testing materials.

Failure on the part of test personnel to follow these testing procedures is a violation of the Code of Maryland Regulations (COMAR) Section 13A.03.04.05A: Test Administration and Data Reporting Policies and Procedures, Testing Behavior Violations.

To further standardize the process during the administration of the MSAs and HSAs, MSDE sends testing monitors, without prior notification, to selected schools to observe administration procedures and testing conditions. All monitors have identification cards for security purposes and follow local procedures for reporting to the school's main office and giving proper notification that an MSDE monitor is in the building.

Test Administration Modes

Select Maryland assessments such as the Mod-HSA and MSA Science are administered in two standardized modes: paper-and-pencil format and a computer administered format. During a computer administration the test items are displayed on a computer monitor (online) in the same order that they appear on a paper-and-pencil test. The test administration is not customized for the student, that is, the items presented are not selected to match the student's ability level as would be the case for a computer adaptive test. Ongoing research is conducted to confirm that the scores resulting from these modes are comparable. How a student performs should not be impacted by the mode of administration taken. For example, the scores from both modes must result in the same performance level classifications and the same proficiency rates at all levels—student, school, school district and disaggregated subgroups.

Scoring Procedures

The Maryland State Department of Education using the competitive RFP process contracts with private companies that specialize in the scoring of constructed response (CR) items to read and score the CRs on the HSA and on the MSA. The readers (also called "scorers" or "raters") work

in sites operated by one of our vendors around the country, such as in Houston, TX; Jacksonville, FL; Atlanta, GA; and Greensboro, NC.

The process begins when the Test/Answer Books are received by the contractor and are scanned into an electronic imaging system so that the information necessary to score responses is captured and converted into an electronic format. Students' identification and demographic information, school information, and answers to SR and SPR (math only) items are converted to alphanumeric format; hand-written responses to CR items are captured in digital image format.

Machine-Scored Items: SRs and SPRs

After students' responses to SR and SPR items are converted to text format, the scoring key is applied to the captured item responses. Correct answers are assigned a score of one point. Incorrect answers, blank responses (omits), and responses with multiple marks are assigned a score of zero.

Hand-Scored Items: BCRs and ECRs

The electronic imaging system, allows readers to score these responses to CRs online at all scoring sites while the live documents are maintained at the contractor's facility. The imaging system randomly distributes responses, ensuring no one scorer scores a disproportionate number of responses from any one school. This online scoring system also maintains a database of actual student responses and the scores associated with those responses. An off-site backup of all images and scores is maintained to guard against potential loss of data and images due to system failure. The system also provides continuous, up-to-date monitoring of all scoring activities.

Highly Qualified Scoring Staff

All staff who work on the Maryland scoring projects must be well educated and highly trained to score using the Maryland scoring criteria.

Readers/Scorers

A Reader/Scorer must have at least a baccalaureate degree from an accredited college or university, have passed a writing sample or other test, have participated in project-specific reader training on Maryland criteria, and have qualified to score using Maryland criteria. The number of readers used depends on how many students take the test and on the deadline for completion of the scoring. Typically, readers score between 30 and 100 CRs per hour of work and work between seven and eight hours a day. The subject area can affect the variance in the rate: reading responses on the MSA take longer to score, on average, than math responses, for example.

Team Leaders and Scoring Directors

A Team Leader (TL) who directly monitors the scoring of a team of readers must be a highly experienced reader who has participated in a two-day training course and qualified to score with at least an 80% perfect match on two of three qualifying sets. A Scoring or Room Director (RD), one for each site, grade and content area, trains the TLs and Readers and supervises the TLs and the scoring of several teams. The RD is typically an exceptional former team leader who has worked on large-scale projects with multiple teams. If any other trainers are used, they are typically former TLs or RDs.

Procedures for Range-finding/Anchor Pulling

Experienced readers sort through a few thousand responses in order to recognize and assemble a wide variety of responses that represent the full range of quality as described in the rubric. They identify not only papers that were homogeneous in their level of quality, but also papers that differ in quality from variable to variable, but which could be given an overall classification of High, Medium, and Low. Readers also identify and flag problem papers—off-topic, off-task, verbatim copying, strange, potential teacher interference, etc. Readers then sort the copies into piles, reflecting the nature of the flag—all potential high papers are together, all potential medium papers are together, etc., with all problem papers grouped together.

When adequate responses are found, the scoring directors come to Maryland, where they meet, once a year, with "rangefinding committees" composed of Maryland educators, MSDE content specialists, and other members of the contractor's scoring staff. These committees discuss in depth the kinds of qualities characterized by certain score points. Papers are read aloud and discussed, a process that focuses attention on what the student had to say—allowing the committee to divorce themselves from how the paper looked or how well it had been edited. Then each member independently assigns an overall tentative score on each response on which there seems to be consensus. An iterative process of reading, charting, and discussing successive sets of papers continues until a firm consensus score is assigned to each paper.

These papers are then used to create training materials that help readers' understand how the responses differ from one another in incremental quality, how each response reflects the description of its score point as generalized in the scoring rubric, and how each reflects MSDE's standard for application of each score point.

Development of Training Materials

Using the officially scored papers from the rangefinding process, scoring directors assemble multiple sets of scored student responses to CR items that are used to train raters and to monitor rater reliability during scoring. These sets, each serving a different purpose, include a "guide" or "anchor" set, "training" sets, "qualifying" sets, and "validity" sets.

Anchor Sets

The anchor sets are composed of 2 to3 clear examples of each score point on the rubric usually organized in ascending or descending order by assigned score. For example, for items scored with a four point rubric, the anchor might contain 12 papers, including two 0's, three 1's, three 2's, two 3's, and two 4's. This anchor set contains only "typical" or "clean" papers, i.e., responses that demonstrate the characteristics of a solid score.

During scoring, readers may compare the response being scored with these anchor papers or student responses from other training materials. For example, after the scanned student's response appears on the computer screen, the reader initially makes a score judgment and then has the option to review anchor papers to confirm or refute this impression. If the reader perceives that the response merits a "1" or a "2," the reader is likely to focus on a few "1" and "2"s in the anchor sets. When the reader identifies a paper that most closely resembles the "live" response being scored, the reader assigns that score.

Training Sets

In contrast, the training sets, composed of about 10 randomly ordered papers each, reflect a broader range of student responses including clean papers, papers with characteristics of two adjacent score points, and papers that show readers unusual, creative, or atypical

approaches to answering the question. The point, as always, is to give readers the most focused training based on the state's criteria. In content areas like English or government, where students' responses may be diverse, it is especially important for readers during training to see as many different student responses as possible.

Qualifying Sets

The qualifying sets are composed of typically 10–20 randomly organized responses per set of mostly clean papers representing the range of rubric score points. They are used following training to ensure that all the readers are competent to score the items they're assigned to score. Readers score the papers in three qualifying tests for each item they will score. A reader who doesn't average a high agreement, approximately 80%, with the official consensus score on the best two sets does not qualify to score actual student responses. Only readers who consistently demonstrate a strong ability to apply the Maryland criteria are allowed to assign scores to Maryland students.

Validity Sets

Validity Sets, also known as calibration sets, are collections of pre-scored responses of mixed quality, arranged in random order. They are regularly circulated, typically without identification, to the scorer during live scoring. The number of validity papers a reader scores during a routine day depends on how fast that reader is scoring. Typically, readers see one validity paper for every 100 actual student responses scored. These sets are designed to ensure that there is no drift in the application of scoring criteria so that scoring remains accurate. Readers who fail to achieve a certain performance level on validity sets are retrained. If after retaining poor performance continues, these readers are dismissed from the scoring process.

Involvement of Maryland Educators

During training and scoring, MSDE content and measurement experts are available to address all questions and concerns. Staff from the Division of Accountability and Assessment are usually at the scoring center itself, and others back in Baltimore from the Division of Instruction are available via phone or email. When a reader, for example, finds a response that doesn't resemble any of the papers in the anchor set, that reader may show the paper to his team leader or the scoring director who was present when the Maryland teachers established how they wanted the item to be scored. Then, if the scoring director can't decide on a correct score, he will send it to MSDE where the best score can be assigned based on careful consideration of criteria Maryland educators have used in the past with similar responses. Thus, scoring decisions can be informed by many experts.

CRs Are Read by (at least) Two Readers

Because it is the best scoring practice, MSDE requires that every CR response be read independently by two readers to ensure the most accurate scoring according to the Maryland guidelines. If both readers give the response the same score, the student receives that score. If two readers assign adjacent scores (such as a "1" and a "2,") the student receives the higher score. If two readers assign non-adjacent scores (such as a "1" and a "3,"), an expert reader, typically the scoring director, assigns a third reader "resolution" score, after consulting the state representatives.

Quality Control

Team leaders, scoring directors, project directors, and MSDE personnel monitor daily reader performance during scoring. They check readers' level of agreement or disagreement with other readers (known as inter-rater reliability), the accuracy of readers' performance on validity sets, and conduct ongoing random "read behinds" (to verify the scores) on a certain percentage of responses. This ensures continued accurate scoring throughout the project. If one reader, for example, has a tendency to assign scores that are lower or higher than other readers in the group, that reader is monitored more closely and retrained.

Equating

MSDE employs Item Response Theory (IRT) analyses as well as classical psychometric techniques to equate tests. Equating adjusts test forms which are designed to be equivalent from a measurement perspective for subtle differences in test difficulty. This adjustment assures the equivalence of assessments and performance standards over time and across different test administrations that use alternate test forms. Equated tests are fair tests for all students and permit accurate comparisons of performance from form to form and from year to year.

There are several methods by which tests may be equated. Maryland currently uses "common item equating" which embeds a core of identical anchor items in each test form. Best practice suggests that each form have a set of at least 20 anchor items or 10% of the items in the entire test, whichever is larger. The anchors are selected taking into consideration essential criteria for technical quality. For example, the content representation of the collection of equating items is similar in overall representation to all of the items in the assessment. The sequence positions of the anchor items are similar to the order in which they were used in previous forms. The distribution of item difficulties for the anchor items is also similar to the overall test and includes the entire range of difficulties, *i.e*. some that are easy, some that are of middle difficulty, and some that are difficult. When possible, all relevant item formats, SRs, BCRs, and SPRs (for math assessments), are represented in the common core of anchor items.

Sampling Procedures

To ensure that the equating sample is representative of the statewide examinee population in terms of gender and ethnicity, a special sampling procedure is used to select the distributions of students by gender and ethnicity.

The ultimate result of equating and scaling is the conversion from raw scores to scale scores to enable the comparison of test scores across test forms for a given content and grade level. Conversion Tables are commonly included in technical manuals.

Statistical Analysis for Items

Maryland uses classical item analyses and item response theory (the application of mathematical models) to provide empirical data for every item on every test form. Typical item statistics include item difficulty, item discrimination, differential item functioning (DIF), and distractor analysis. Each is used to judge whether a given question is suitable for inclusion in the item bank from which operational forms will be assembled. Item analysis data and the expert judgments of content and measurement specialists permit the detection of flaws before the items are used on operational tests. Items which can be revised are rewritten and re-field tested. Any items that are fatally flawed are discarded.

Item Difficulty

Item Difficulty is the percent of test-takers who answer an item correctly. This value should be appropriate for the intended test takers. When an item is too easy, virtually all test takers answer it correctly; thus, extremely easy items contribute very little information to the total test score. Similarly, inappropriately difficult items are not very useful in a test. Items with a moderate spread of difficulty around a mean permit greater item discrimination.

Item Discrimination

Item Discrimination is an indication of the relationship between performance on an item and on the overall test, or how well an item differentiates between low and high performers. Students who perform well on the test should get the item correct; conversely, students who perform poorly should get the item incorrect. When the reverse is true, the item discriminates poorly, or not at all, and the statistic suggests that the item is flawed.

Differential Item Functioning

Differential Item Functioning (DIF) analyses are conducted to identify items that may function differently for members of different groups. DIF analyses compare the performance of two groups of test-takers (e.g., males vs. females, African American test takers vs. white test takers) who have been matched on their proficiency as measured by the test. The underlying assumption in conducting such analyses is that all test-takers demonstrating the same level of proficiency in a subject should have similar chances of answering each item correctly, regardless of gender, race, or ethnicity. When comparably proficient groups do not perform similarly, the statistics suggest that the item may be biased and must be revised or discarded.

Distractor Analysis

Distractor Analysis looks at the frequency of the selection of the incorrect answer choices for SR items called "distractors" They are called distractors because they are intended to attract the uninformed or partially informed student from the correct answer. A distractor that is selected more frequently than the intended correct answer suggests that the item may have more than one defensible correct answer when only one answer may be selected, or conversely, may have no defensible correct answer. A distractor which is never or infrequently selected is implausible, and hence because it is nonfunctional, it contributes nothing to the measure process. It also allows the less informed student to more easily guess the answer without having the knowledge the item is intended to measure.

Standard Setting

Maryland, like most states, uses a carefully managed standard setting process, known as item-mapping or the Bookmark technique to identify proficiency levels and cut scores. There are sometimes minor variations in the standard setting procedures necessary to accommodate a specific test.

The process has resulted in a long history of successful standard setting. The methodology is nationally accepted by prominent measurement experts and most state departments of education and has been approved and validated by Maryland's National Psychometric Council and by national peer reviewers for the United States Department of Education (USED). It is a careful and well-planned process involving content experts and other educational and business stakeholders. Each standard setting is conducted by a contractor according to MSDE specifications that are

reviewed and approved by the National Psychometric Council prior to implementation and validated after implementation. Ultimately, all state procedures for standard setting and results for assessments required by No Child Left Behind legislation are subjected to peer review by the USED. Procedures that do not satisfy federal requirements will result in disapproval of a state's accountability system.

Maryland's standard setting process includes five steps, each involving a different group of experts and stakeholders. Each group makes recommendations which are relayed to the next group. All groups clearly understand that they are making recommendations that will finally be acted upon by the Maryland State Board of Education.

Content Reviewers

The first step involves the input of content experts specific to the grade levels assessed representing many areas of the state. These are groups of teachers and some local central office staff with expertise in the content area and a strong familiarity with the Voluntary State Curriculum, as well as the specific students assessed by the test. They receive training on the standard setting process, and then begin activities to familiarize themselves with the specific content and grade level assessment with which they are working, samples of student responses, the proficiency level descriptors for student performance and other information germane to their task. Once they are thoroughly prepared, they begin the Bookmarking Standard Setting Procedure.

This procedure is used throughout the nation for establishing achievement standards in other state level and commercially produced assessment programs. The procedure requires the experts to work individually and then in discussion groups to come to an agreement on the lowest performance that can be defined as *proficient*, and then the lowest performance that can be considered *advanced*. They complete two rounds of standard setting decisions and discussions, and then individually determine their own final recommendation. The median (midpoint) of the groups' final recommendations becomes the cut score (standard) that is forwarded to the next group.

To ensure quality, each panelist has three opportunities to evaluate the standard setting procedures: 1) readiness surveys are completed prior to each round of judgments to determine whether training activities have prepared panelists for their tasks, 2) questionnaires, completed post all rounds of judgment, seek panelists' level of comfort with the procedure, their understanding of the performance levels, and their satisfaction with the final cut score; and finally 3) exit surveys evaluate the overall process.

Psychometric Review

The integrity with which the process is conducted is of paramount importance. The next step is a thorough review of the process followed by Maryland's National Psychometric Council, an independent nationally recognized panel of testing experts. Their charge is to determine that the standard setting process was a valid one. They also provide guidance that assists in the next steps of the process to ensure that any adjustments made by future groups to the content team recommendations are appropriate. Maryland's National Psychometric Council guards carefully the quality of the process by assuring that all steps of the process are properly done. However, it is not the role of the council to change the recommendations for standards, but rather to assure the integrity of the process.

Review and Articulation

The review and articulation panel is broadly representative of school system leaders, educators, and stakeholders (including parents and other advocates, and business representatives). Their charge is to make a recommendation to the State Superintendent of Schools based on their review of the recommendations of the content groups and the Psychometric Council. They examine the consistency of the recommendations within a content area, across grade levels and across content areas.

State Superintendent's Review

MSDE staff meets with the State Superintendent of Schools to assist her in a review of the recommendations of the content teams, the Psychometric Council, and the review and articulation panel. The State Superintendent examines the recommendations within the context of other state assessments and national assessments. The recommendations of the State Superintendent of Schools for standards reflect the composite of information accumulated during the entire process of study and deliberations.

State Board Approval

The State Superintendent takes the recommendation to the State board of Education for final approval.

Implementation

Once standards have been approved they are applied to student scores, verified, validated and released to local school systems. School systems are responsible for distributing results to schools and parents.

Performance Level Standards and Descriptors

The implementation of the standard setting process, yields three levels of academic achievement standards (Basic, Proficient, and Advanced) and the cut scores for these levels against which yearly results are compared.

Advanced is a highly challenging and exemplary level of achievement indicating outstanding accomplishment in meeting the needs of students.

Proficient is a realistic and rigorous level of achievement indicating proficiency in meeting the needs of students.

Basic is a level of achievement indicating that more work is needed to attain proficiency in meeting the needs of students.

For each content area and grade assessed by the MSAs and HSAs, each proficiency level includes sample **Performance Level Descriptors (PLDs)** that indicate the general knowledge and skills students have when their test scores fall within that level. These descriptors are directly linked to the indicators and assessment limits that the test items measure. Table 3 shows Selected Proficiency Level Descriptors for the "Proficient" Standard for the Algebra/Data Analysis HSA/MSA. The complete lists of PLDs for each assessed content and grade level are posted on the web.

Table 3: Selected Proficiency Level Descriptors for the "Proficient" Standard for the Algebra/Data Analysis HSA/MSA

Proficient Standard for Algebra/Data Analysis	
What *proficient* students likely can do that *basic* students cannot do: • Recognize how scales of a graph can lead to misuse of data • Identify the graph of a system of equations of the form y=mx+b • Extend a linear pattern to terms beyond the first several given terms, explain the process • Write the equation for a line of best fit • Determine the quartiles of a data set • Identify and justify that a representative sampling method provides variety in the sample • Selecta simple random sampling method from a list of sampling methods • Use a curve of best fit to describe the trend of the data and justify when it is inappropriate use the curve to make a prediction • Identify a line of best fit that models the data on a scatterplot and use the graph to make a prediction • Determine the difference between two matrices	**What *proficient* students likely cannot do:** • Determine the range of a non-linear gragh • Write an inequality in the form ax+by ≥c that models a real-world situation from a verbal description • Extrapolate the value of a graph beyond the grid provided • Fully justify why a sampling method will or will not provide a representative sample • Write and solve a system of equations in the form ax+by=c that models a real-world situation from a verbal description as well as explain and justify the process • Extend a linear pattern to terms beyond the first several given terms, explain and justify the process • Use a curve of best fit to describe the trend of the data, to make a prediction close to the data set and to justify when it is inappropriate use the curve to make a prediction

Reporting

Maryland's assessment reports represent the culmination of all other aspects of its standards and assessment system. In these reports, a parent, educator, or other stakeholder can find answers to questions about how well a student or group of students is achieving, as well as important information on how to improve achievement in the future.

Maryland produces reports at the individual student, school, school district, and State levels. At each of these levels, reports include scores that are aligned with the Maryland's voluntary state curriculum. Also, total test scores are reported by performance levels—Basic, Proficient and Advanced—defined in the State's academic achievement standards.

The test development contractor verifies the accuracy of the data using sophisticated computer programs that scan for irregularities. If none are found, each score report can be produced. Every effort is made to disseminate the reports as soon as possible after each assessment administration. The individual student reports are also accompanied by score interpretation guides that help parents and educators make appropriate, credible, and defensible score interpretations and use the information the reports provide for school improvement.

Maryland carefully protects the data files containing student-level information that are produced following each assessment administration. When the State allows access to this information, it does so in a way that maintains the confidentiality of each student's records.

Assessment of English Language Learners (ELL)

Students who have been identified for participation in a language instruction educational program are tested for their knowledge of English using the Language Assessment Scales (LAS) Links. Based on this standardized assessment they are designated as (1) Beginner; (2) Intermediate; (3) Advanced; or (4) Proficient. Those identified as Beginner have no or very minimal English Language Proficiency.

Students in their first year of enrollment in U.S. schools can use the LAS links in place of the MSA reading assessment to meet their testing requirement. However, all ELL students must take the MSA mathematics assessment, although the scores do not count in school accountability. ELL students have an accommodations plan to support them during instruction and many of these accommodations are also available to them during state assessments.

Assessment of Students With Special Needs

In Maryland, students with disabilities participate in the assessment type appropriate to their needs. They may take the Maryland School Assessment (MSA) in reading, mathematics, and science with or without accommodations, as appropriate. They may take the Alternate Maryland School Assessment based on Alternate Academic Achievement Standards (Alt-MSA) in reading, mathematics, and science if they have an Individualized Education Program (IEP). Or they may take the Alternate Maryland School Assessment based on Modified Academic Achievement Standards (Mod-MSA) if a regular grade-level MSA is too challenging and the Alt-MSA is not appropriate. The participation guidelines for Mod-MSA appear in the 2007-2008 Maryland Accommodations Manual. The MSA and Alt-MSA are administered to students in grades 3−8 in reading and mathematics and in grades 5 and 8 in science.

At the high school level students with disabilities also participate in the end-of-course assessment type appropriate to their needs. They may take the High School Assessments (HSAs) in English 2, algebra/data analysis, government, and biology (with or without accommodations, as appropriate) when they complete the course. They may take the Mod-HSAs in English 2, algebra/data analysis, government, and biology or the Alt-MSA in reading, mathematics, and biology in grade 10. All of the assessments are meticulously aligned with the Maryland Voluntary State Curriculum (VSC) Content Standards.

Alternate Maryland School Assessment (Alt-MSA)

The Alt-MSA is the Maryland assessment in which students with significant cognitive disabilities participate, if through the IEP process it has been determined they cannot participate

in the Maryland State Assessment even with accommodations. Alt-MSA measures a student's progress on attainment of Mastery Objectives in reading and mathematics in grades 3 through 8 and 10 and in science for students in grades 5, 8, and 10. A portfolio is constructed of artifacts that document individual student growth in the assessed objectives. An artifact is a sample of student work such as a data collection chart, videotape, or audiotape that shows that the student has mastered the objective. Students' proficiency levels are included in districts' AYP reports.

Modified High School Assessment (Mod-HSA)

The Mod-HSA is the alternate assessment with modified achievement standards in reading, mathematics, and science for grades 5, 8, and 10-12. This assessment is taken only by the small group of students with disabilities who can make significant progress, but who may not reach grade-level achievement in the time frame covered by their IEPs.

Accommodations

It is important to note that when students take an assessment with approved accommodations that the accommodation does not alter what is being measured, and thus does not invalidate a student's score. Typical acceptable accommodations include changes to the presentation format, the response format, test scheduling and timing, and test setting. Readers interested in a Quick Reference Guide to Accommodations for students with Disabilities (SWD) and a Quick Reference Guide to Accommodations for English Language Learners (ELLs) should consult appendices C and D, respectively.

Maryland K–12 educational stakeholders interested in detail about assessment design and development beyond the scope of this paper should consult the references and the glossary.

References

A Continuum of Quality. Designing Assessment Systems: A Primer on the Test Development Process (2003). Monterey: CA www.CTB.com Available at the link: http://www.ctb.com

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: AERA.

Brennan, R. L. (Ed.) (2006). Educational measurement, 4th edition. Ace/Praeger Series on higher education. National Council on Measurement in Education. Portsmouth, NH: Greenwood Publishing Group, Inc.

CCSSO SCASS TILSA (2003). Quality control checklist for item development and test form construction. Washington, DC: Author. http://www.ccsso.org/content/pdfs/ItemandTestDevQCChklst.pdf

Code of Fair Testing Practices in Education. (2004). Washington, DC: Joint Committee on Testing Practices. http://www.ncme.org/pubs/pdf/CodeofFairTestingPractices.pdf

Code of Professional Responsibilities in Educational Measurement. (1995). NCME Ad Hoc Committee on the Development of a Code of Ethics: Washington, DC: National Council on Measurement in Education. http://www.natd.org/Code_of_Professional_Responsibilities.html

Cizek, G. (Ed.) (2001). Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

Downing, S.M. & T.M. Haladyna (Eds.) (2006). Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

ETS Fairness Review Guidelines (2003). Princeton, NJ: Educational Testing Service.

ETS Standards for Quality and Fairness (2002). Princeton, NJ: Educational Testing Service.

Fairness Report for the ACT Tests 2005-2006 (2006). Iowa City, IA: American College Testing, Inc.

Johnstone, C. J. (2003). Improving validity of large-scale tests: Universal design and student performance (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. http://education.umn.edu/NCEO/OnlinePubs/Technical37.htm

Maryland Assessment Manual: Selecting, administering and evaluating the use of accommodations for instruction and assessment. (February 15, 2008). Baltimore, MD: The Maryland State Department of Education.

Maryland High School Assessment Technical Report for Algebra and Data Analysis, Biology, English, Geometry, Government. (2005) Princeton, NJ: Educational Testing Service.

Maryland School Assessment (MSA) Science, Grades 5 and 8, Technical Report 2007 Field Test. (February, 2008). Iowa City, Iowa: NCS Pearson.

Modified Academic Achievement Standards. (July, 2007). Non-Regulatory Guidance. Washington, DC: U. S. Department of Education Office of Elementary and Secondary Education.

National Council of Teachers of Mathematics (1996). Principles and Standards for School Mathematics. www.NCTM.org/standards

National Science Education Standards. (1996). National Research Council. Washington, DC: National Academy Press. http://www.nsta.org/publications/nses.aspx

National Standards for Civics and Government. (1994-2007). http://www.civiced.org/index.php?page=stds_toc_intro

NCTE/IRA Standards for the English Language Arts. (1996). International Reading Association and the National Council of Teachers of English. http://www.ncte.org/about/over/standards.

Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001, (Revised December 21, 2007). Modified academic achievement standards. Washington, DC: U. S. Department of Education Office of Elementary and Secondary Education.

Appendix A

Glossary

Academic Achievement Standards *(See also "Performance Level Standards.")*

- A grade-level academic achievement standard (GLA) defines a level of "proficient" performance equivalent to grade-level achievement on the State's regular assessment.
- An alternate academic achievement standard (AAS) is an expectation of performance that differs in complexity from a grade-level achievement standard, usually based on a very limited sample of content that is linked to, but does not fully represent grade-level content. AASs were authorized under USED regulations (34 C.F.R. Part 200) published on December 9, 2003.
- A modified academic achievement standard (MAS) is aligned to grade-level content standards for the grade in which a student is enrolled and challenging for eligible students, but may be less difficult than grade-level achievement standards. In Maryland the MAS and the GLE standards for the high school assessments are the same. MASs were authorized under USED regulations (34 C.F.R. 200 and 300) published on April 9, 2007.

Academic Content Standards Challenging, but fair, grade-level expectations of what students should know and be able to do. See the Core learning Goals; the Voluntary State Curriculum

Accommodation A change in the way an assessment is given or taken that does not alter what is being measured, and thus does not invalidate a student's score. Students with IEPs or 504 plans who are receiving accommodations in instruction are eligible to have those accommodations during testing. Typical acceptable accommodations include changes to the presentation format, the response format, test timing and test setting. See the Maryland Accommodations Manual (February 15, 2008) for complete information.

Adequate Yearly Progress (AYP) AYP is the gain that schools, school systems, and states must make each year in the proportion of students achieving proficiency in reading and math toward the NCLB goal of 100% proficiency in 2014. AYP replaces the School Performance Index as the method by which Maryland tracks academic progress and makes accountability decisions. To make AYP, schools and school systems must meet three criteria: 1) the annual measurable objective (AMO) in reading and mathematics for students in the aggregate and for each student subgroup, 2) the graduation rate for high school or attendance in elementary and middle school for students in the aggregate, and finally 3) the testing participation requirement of 95%. See Differentiated Accountability

Advanced Placement Assessments (AP) College-level AP courses and exams designed by the College Board give students the opportunity to earn credit or advanced standing at most of the nation's colleges and universities. Maryland permits students who take select AP courses to substitute performance on select AP Assessments for Maryland's High School Assessments in grade 10 English, Algebra/Data Analysis, and Geometry. Maryland used decision classification consistency to identify the minimum score on an AP Assessment (≥ 3 on a scale of 1 to 5) that a student must achieve in order to met or exceed the "Proficient" academic achievement level (passing score) on the respective designated Maryland high school assessment.

Aggregated Data Score results based on all test takers.

Alert Forms Forms completed by readers during the scoring of constructed response items used to report suspected teacher interference, possible student suicide threats, or any other evidence of student distress. Once notified MSDE, as appropriate, follows up with the State Test Security Committee and/or the local school systems.

Alternate Forms A generic term for two or more test forms that are considered to be interchangeable because they were designed for the same purpose to measure the same learning outcomes using the same item formats in the same proportion and administered by the same standardized procedures. Examples of alternate forms which have different statistical properties are parallel forms, equivalent forms, and comparable forms. Information on form-to-form score equivalence is contained in the assessment's technical manual.

Alignment A judgment made by trained subject matter experts of whether the test items on an operational assessment match the academic content standards the assessment intends to measure. Several methods of evaluating alignment between standards and assessments have been developed. A summary and comparison of alignment models is found on the Council of Chief State Officers website at: http://www.ccsso.org/Projects/alignment_analysis/models/418.cfm

Alignment Criteria The four criteria which are typically examined were proposed by Norman Webb: Categorical Concurrence, Depth of Knowledge Consistency, Range of Knowledge Correspondence, and Balance of Representation. See the alignment tool at http://wat.wceruw.org/index.aspx.

- **Categorical Concurrence** is used to judge whether the assessment includes items measuring content from each of the expectations which are used for sub-scores.
- **Depth of Knowledge Consistency** is used to determine whether the assessment items are as demanding cognitively as what the students are expected to know and do as stated in the expectation.
- **Range of Knowledge Correspondence** is used to judge whether the span of knowledge expected by an expectation is the same as, or is comparable to, the span of knowledge that students need in order to answer the assessment items.
- **Balance of Representation** is used to determine the percent of indicators within each expectation that the assessment addresses.
- **Source-of-Challenge** A Webb alignment criterion that is only used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted skill, concept, or application. Cultural bias or specialized knowledge could be reasons for an item to have a source-of-challenge problem. Such item characteristics may result in some students not answering that item, or answering it incorrectly or at a lower level, even though they possess the understanding and skills being assessed.

Alternate Maryland School Assessment (Alt-MSA) The Maryland assessment in which students with significant cognitive disabilities participate if through the IEP process it has been determined they cannot participate in the Maryland State Assessment even with accommodations. Alt-MSA measures a student's progress on attainment of Mastery Objectives in reading and mathematics in grades 3 through 8 and 10 and in science for students in grades 5, 8, and 10. A portfolio is constructed of artifacts that document individual student growth in the assessed objectives. An artifact is a sample of student work such as a data collection chart, videotape, or

audiotape that shows that the student has mastered the objective. Students' proficiency levels are included in districts' AYP reports.

Alt-MSA Performance Level Descriptors Like the grade-level MSA, performance is reported using three Performance Standards: Advanced, Proficient and Basic.

- **Advanced** students demonstrate 90% or greater attainment of their identified mastery objectives in reading and mathematics (attainment of 9 or 10 of the student's Mastery Objectives in a given content area).
- **Proficient** students demonstrate at least 60% but less than 90% attainment of their identified mastery objectives in reading and mathematics (attainment of 6 to 8 of the student's Mastery Objectives in a given content area). The goal for all students is to reach the proficient or advanced level.
- **Basic** students demonstrate 0% to less than 60% attainment of their identified mastery objectives in reading and mathematics. (attainment of up to 5 of the student's Mastery Objectives in a given content area).

Anchor Items Anchor items are test items that are embedded within each of the alternate forms for a content area and grade that permit these forms to be linked (equated) within grade level and across years to determine the comparability of the results.

Anchor Papers To score constructed response items, readers must be trained to score student responses according to the criteria at each score point on a scoring tool called a rubric. Anchor papers are actual samples of student work that best represent each score point. They help to ensure the consistency (inter-rater reliability) with which readers assign scores.

Annual Measurable Objectives (AMO) State established performance targets that assess the progress of student subgroups, schools, school districts, and the state annually. This annual measurement ensures that 100% of students achieve proficiency in reading/language arts and mathematics by the end of the school year in 2013–14.

Between the 2002–03 baseline and the 2013–14 goal of 100% proficiency, Maryland established annual measurable objectives for reading, mathematics, attendance, and graduation rate. Every school and school system is held to the same annual measurable objectives, although those objectives are adjusted to each school's grade-level enrollment and structure (e.g., K-5, 6-8, K-8, K-12). Schools with grade structures that do not include tested grades will still be accountable for student performance; e.g., the performance of third-graders who come from K-2 schools will count for both the current school and the K-2 school previously attended.

Artificial Intelligence (AI) Scoring A process by which AI is used to score student responses to constructed response items. MSDE is researching the feasibility of implementing an AI system as the *second* read for constructed response items as a potential future source of cost-containment and improved test score turn-around time.

Assessment Limits The objectives assessed on MSA at each grade level, called assessment limits, are embedded in the Core Learning Goals (CLG). They identify the specific skills and content that students are expected to have learned for each assessed objective. Not all objectives in the VSC have Assessment limits at a given grade-level. Nonetheless, these objectives must be included in instruction because they introduce important concepts in preparation for assessed skills and content at subsequent grades.

Bias in Testing Deficiencies in a test or the manner in which it is used that result in different score interpretations for different subgroups defined by race, ethnicity, gender, and disability.

Bridge Plan for Academic Validation A student who has failed an HSA twice and meets eligibility criteria, including locally-administered or approved assistance, can complete one or more project modules in the content area failed. The project modules will be submitted to a local review panel and the local superintendent for approval. http://www.hsaexam.org/about/options/bridge_plan.html

Council of Chief State School Officers (CCSSO) A nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

Classroom Assessments Teacher designed and developed tests and quizzes often produced at the teacher or school level by teams of collaborating subject matter experts that help them tailor their lessons to students' strengths and weaknesses. In this way, good tests provide teachers with insight into their own methods and feedback on their teaching practice.

Computer Administered Test (CAT) A test administration in which the test items are displayed on a computer monitor (on line) rather than in paper-and-pencil format. The test administration is not customized for the student. Maryland has online tests available for science, modified assessments, and beginning in May 2009, High School Assessments.

Concurrent Validity The degree to which a test designed to measure specific content (3rd grade math skills) correlates with other measures of the same content (3rd grade math skills) measured at the same time.

Construct Irrelevance Also known as construct over-representation, construct irrelevance is the presence of test items on an assessment that measure knowledge, abilities, and skills beyond those that the assessment intends to measure (i.e., the assessment limits). Such an assessment is not “aligned” with the academic content standards.

Construct Underrepresentation A condition that results when the sample of test items on an assessment does not measure all of the knowledge, abilities, and skills that the assessment intends to measure (i.e. higher-order thinking skills). Such an assessment is not “aligned” with the academic content standards.

Construct Validity The degree to which a test designed to measure a specific construct (i.e., theoretical trait) measures that trait or behavior. Four broad categories of evidence are used to determine construct validity: (1) evidence based on test content, (2) evidence based on the assessment's relation to other variables, (3) evidence based on student response processes, and (4) evidence from internal structure.

Content Validity A judgmentally determined validity that involves "the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured. See alignment.

Convergent Validity An empirical validity that describes the degree to which a measure is correlated with other measures with which it is theoretically predicted to be related.

Conversion Tables The ultimate result of equating and scaling is the conversion from raw scores to scale scores to enable the comparison of test scores across test forms for a given content and grade level. Conversion Tables are commonly included in technical manuals.

Core Learning Goals (CLGs) Grade-level goals, expectations and indicators that define what a student should know and be able to do in English, Government, Algebra/Data Analysis, Biology, Geometry, Physics, Chemistry, Earth/Space Science and Environmental Science. The goal statements in English, Government, Algebra/Data Analysis, and Biology also contain assessment limits because these content areas are the basis for the four high school assessments students must pass to graduate beginning in 2009 (http://www.mdk12.org/assessments/standards/9-12.html).

Criterion-referenced Test (CRT) Interpretation CRTs, referred to as standards-based tests or proficiency tests, measure how well a student measures up to a certain criterion or standard. Scores indicate how close the test-taker is to meeting the standard in a given subject —for example, doing eighth-grade math—regardless of how well other eighth-grade math students performed on the test. Technically an assessment can yield both criterion-referenced *and* norm-referenced interpretations, provided this is the goal of the test design. Thus, it is not the instrument itself that is an NRT or a CRT, per se, but the interpretation attached to the scores. The design and intended use of the scores impacts the interpretation that may be attributed to them.

Criterion-related Validity This validity reflects the success of measures used for prediction or estimation. There are two types of criterion-related validity: Concurrent and predictive validity. A good example of criterion-related validity is in the validation of employee selection tests; in this case scores on a test or battery of tests is correlated with employee performance scores.

Cut Score Scores for each content area and grade measured by the Maryland assessments that are determined during the standard setting process that define the three proficiency level categories of Basic, Proficient and Advanced. Cut score are approved by the State Board of Education The cut score that defines "proficient" for the HSAs is also called the "passing score." Cut scores for all content areas and grades tested are shown on the web at www.HSAexam.org.

Data Analysis Learn how to analyze your state assessment data at http://mdk12.org/data/.

Decision Accuracy The extent to which one decision would agree with the decisions that would be made if each student could somehow be tested with all possible parallel forms of the assessments.

Decision Consistency The extent to which one decision would agree with the decisions that would be made if the students had taken a different form of the examination, equal in difficulty and covering the same content as the form they actually took

Decision Misclassification There are two types of misclassifications: false positives and false negatives. Students who were below the proficiency cut score, but were classified (on the basis of the assessment) as being above a cut score, are considered to be false positives. Students who were above the proficiency cut score, but were classified as being below a cut score, are considered to be false negatives

Differential Item Functioning (DIF) Analyses of DIF are conducted to identify items that may function differently for members of different groups. DIF analyses compare the performance of two groups of test-takers (e.g., males vs. females, Asian American test takers vs. white test takers) who have been matched on their reading, writing, or mathematical proficiency. The underlying assumption in conducting such analyses is that all test-takers demonstrating the same level of proficiency in the content area should have similar chances of answering each item correctly, regardless of gender, race, or ethnicity.

Differentiated Accountability Maryland is one of six states (Florida, Georgia, Illinois, Indiana, and Ohio) approved for a federal accountability pilot. Maryland's proposal, implemented in the 2008–09 school year, categorizes schools in need of improvement in two ways: the Comprehensive Needs pathway and the Focused Needs pathway. Comprehensive Needs schools are those that do not make their progress targets in the "all students" category, or are having difficulty making targets in three or more subgroups. Focused Needs schools have achieved targets in the "all students" category but have not achieved targets in one or two subgroup areas.

Discriminant Validity An empirical validity that describes the degree to which a test does not correlate with other with other tests with which it theoretically should not be related.

Distractor Analysis A statistical analysis that looks at student performance on each of the options on a multiple-choice test to determine whether there are anomalies in performance, such as when higher achieving students get an item wrong while low-performing students get it correct. It can also identify implausible options that are not adding any information to the overall test item.

Disaggregated Data Data that are reported for special populations.

Documentation Complete, current, accurate and clear information written by test developers for diverse educational stakeholders such as other measurement experts, parents, students, teachers, and administrators. Documents typically specify the nature of the assessments, their intended use, content and skills that are measured, the procedures used for the development, administration, scoring, reporting, and interpreting the results. Some are secure (e.g., test specifications, test items, operational test forms); most are non secure (e.g., the core learning goals, the voluntary state curriculum, handbooks for parents and students, test administration manuals, scoring guides, score interpretation manuals, state and school district score reports, annual disclosed test forms, technical manuals, research studies, sample items of the type included on the assessments, characteristics of educators who participated in item reviews, standard setting, and alignment studies).

Educational Taxonomy Cognitive domains (i.e., thinking skills) important for learning, teaching and assessing. Popular taxonomies include those of Bloom, Marzano, and Wiggins & McTighe. *(See "Understanding By Design.")*

English Language Learners Students who have a primary or home language other than English and who may have limited or no age-appropriate ability to understand, speak, read, or write English. ELL students have traditionally been known as Limited English Proficient (LEP) students and are served with English for Speakers of Other Languages (ESOL) classes or services.

Equating A statistical procedure for placing two or more tests that measure the same traits on a common scale.

Face Validity Closely related to content validity, face validity is made by a cursory inspection of the test items); thus it can be judged by the amateur. It relates to whether a test appears to be a good measure or not.

Fairness in Testing In the measurement context, fairness typically has four criteria: 1) absence of bias, 2) equitable treatment of test takers before, during and after testing, 3) equity in opportunity to learn the knowledge and skills measured by the test, and 4) tests composed of items which test takers of comparable ability should be able to answer correctly [items free of DIF]. Fairness does not require overall passing rates across subgroups. Fairness is enhanced by providing students who do not pass with multiple opportunities to retake alternate forms of the test, ensuring that students have prior knowledge of the content and skills that will be tested and test preparation materials, providing students with an adequate opportunity to learn the measured knowledge and skills, and proving students with test taking strategies.

Field Test A test administration composed of only items that have not previously been pre-tested used to monitor the adequacy of the test development process. Occasionally, a field test also serves as an "operational" test in which the scores count and the item data are used to revise the items for subsequent operational test forms. The spring 2007 administration of the science MSA was an "operational" field test administration.

Field Test Items Clusters of FT items may be administered "standalone" or embedded among operational items. Standalone field tests are composed of only items that are included for the purpose of collecting item analysis information that can be used to review and revise these items prior to using them on tests which count towards a student's score. Embedded FT items are items that do not count towards a student's score, but are distributed throughout the operational test so that students are motivated to respond optimally to them as if they thought they were going to count. They, too, are included for the purpose of collecting item analysis information for item improvement.

Free and Reduced Price Meals (FARMS) Students whose applications meet family size and income guidelines of the United States Department of Agriculture.

Hand-scored Tests Hand-scored tests are scored manually rather than by machine, and require human judgment in the scoring. Examples of hand-scored tests include authentic measurement tests, constructed-response tests, essay tests, and performance tests.

High School Assessments (HSAs) The High School Assessments are high-stakes, end-of-course, standards-based tests in English2, government, algebra/data analysis, and biology that all students who take specific high school level courses must take upon completion of the respective courses. Separate student scores for each test are reported for the state, school systems, and

schools. The passing scale scores for all four of the content areas are posted on the web at www.HSAexam.org. High School Graduation Requirements Questions & Answers, a detailed document, is on the web at www.marylandpublicschools.org.

High-Stakes Test A test used to provide results that have important, direct consequences for students, programs, schools, and school systems. Examples include tests that must be passed for a student to graduate from high school and tests used for NCLB accountability.

Highest Obtainable Scale Score (HOSS) As the name states, this score is the highest score possible. For the MSAs with a scale ranging from 240 to 650, the HOSS is 650.

Home Schooled Students (LEA 55) Students who receive their instruction in their own homes and who are permitted, at the discretion of their parents/guardians, to be included in the testing at a public school location. Scores from home-schooled students are not included in any way in the state accountability system. These students are distinct from "Home and Hospital" students, who are officially enrolled in a public school system and who *are* included in the state accountability system.

International Baccalaureate Assessment (IB) The IB Diploma Program is a challenging two-year curriculum, primarily aimed at students aged 16 to 19 that leads to a qualification that is widely recognized by the world's leading universities. Maryland permits students to substitute performance on select IB Assessments for Maryland's High School Assessments in grade 10 English, Algebra/Data Analysis, and Geometry. Maryland used decision classification consistency to identify the minimum score on an IB Assessment ($\geq$ 5 on a scale of 1 to 7) that a student must achieve in order to meet or exceed the "Proficient" academic achievement level (passing score) on the respective designated Maryland high school assessment.

Inter-rater Reliability The degree to which two raters reading open-ended responses assign the same score to the same paper. Agreement may be "exact"-two raters assign the same score (1, 1); adjacent—two raters assign scores within one score point of each other (1 and 2, or 2 and 3)—or non-adjacent—two raters assign scores within two score points of each other (1 and 3, or 2 and 4). Non-adjacent scores require a third reader resolution. High inter-rater agreement does not guarantee high reliability of total test scores.

Item Banks Item banks, also known as item pools, are collections of hundreds of field tested high quality test items which are available for inclusion on test forms. The test contractor maintains a computerized statistical item bank to store supporting and identification information on each item. The information stored in this item bank for each item may include: the current item ID, test administration year and season, test form, grade level, item type, item stem and options for SRs, passage code and title, subject code and description, process code and description, standard code and description, indicator code and description, objective code and description, item status, and item statistics.

Item Characteristic Curve (ICC) It shows the probability of a correct response to an item as a function of the ability level. The probability of a correct response is bounded by 1 (certainty of a correct response) and 0 (certainty of an incorrect response).

Item Difficulty In classical test theory, the item difficulty is the percent of test-takers who answer an item correctly (called the *p*-value). This value should be appropriate for the intended

test takers. When an item is too easy, virtually all test takers answer it correctly; thus, extremely easy items contribute very little information to the total test score. Similarly, inappropriately difficult items are not very useful in a test. Because items within a test are highly inter-related, it is best to select items with a moderate spread of difficulty around a mean (average) *p*-value of .5 (or 50% correct).

Item Discrimination One statistic used to determine item discrimination is called the "point biserial (R-bis)" and is the relationship between performance on an item and on the overall test, or how well an item differentiates between low and high performers. R-bis values range from -1.0 to 1.0 (similar to a correlation coefficient), with values above .30 meaning the item discriminates well, values close to zero meaning the item discriminates poorly, and values that are negative meaning that the item is most probably flawed.

Item Formats Item formats used on the Maryland School Assessments include: Selected Response (SR), also known as Multiple-Choice (MC); Constructed Response (CRs) which includes Brief Constructed Response (BCR), Extended Constructed Response (ECR); and Student-Produced Response (SPR).

- **Selected Response items (SRs)** pose a question followed by suggested answers only one of which is correct or clearly best. SR items on grade-level tests typically have four or five answer options per item; SR items on the alternate assessments with modified achievement standards (AA-MAS) called the "Mod-HSAs" have three answer choices per item. SRs are either stand alone (not linked to other items), linked to a stimulus set (e.g., technical graphs or figures of lab sets), or linked to a passage stimulus.) Students record their answers to SRs on a machine-scorable grid call a "bubble sheet." SRs are typically scored right (1 point) or wrong (0 points).
- **Constructed response items (CRs)** include BCRs and ECRs. Each asks a question (called a "prompt") to which the student must compose a short or extended answer, respectively. Like the SRs, they, too, may be stand alone or linked to stimulus material. On the math MSAs, both the BCR and the ECR consist of a Step A and a Step B. Step A contributes to the content score and Step B contributes to the process score. All BCR and ECR Step A items receive a 0–1 score point range from two independent scorers; all BCR Step B items receive a 0–2 score point range and all ECR Step B items receive a 0-3 score point range from two independent scorers. On the reading MSAs, the BCRs are scored with a general rubric with maximum values between 0–3.
- **Student-Produced Response (SPR)** Used only on the math assessments, these items require students to calculate and record an answer on a machine-scorable grid. Answers may be whole numbers, fractions, or decimals. SPRs are typically scored right (1 point) or wrong (0 points).

Item Response Theory (IRT) Often referred to as latent trait theory or strong true score theory, IRT is a body of theory describing the application of mathematical models to data from surveys and tests as a basis for measuring abilities, attitudes, or other variables. It is used for statistical analysis and development of assessments, often for high stakes tests such as those required for graduation. At its most basic level, it is based on the probability of getting a test question correct as a function of a latent trait or ability. For example, a student with higher knowledge and skills in 5th grade math would be more likely to correctly respond to a given item on a test of 5th grade math. Among other things, IRT theory provides a basis for evaluating how well assessments and individual questions on assessments work.

Item Statistics Classical item analyses involve computing, for every item on every test form, item statistics such as item difficulty, item discrimination, and differential item functioning. These statistics are used to judge whether a given question is suitable for inclusion in the pool of items from which operational forms will be assembled. The item statistics may also reveal problems with the conceptualization or wording of a question. Some of these items will be revised and re-pretested. Others will be discarded.

Item Weighting *(See "Pattern scoring.")*

Kurzweil™ Verbatim Reading Materials Students with an Individualized Education Plan (IEP) or other plan which specifies a verbatim reading presentation accommodation may have the test read to them by a reading software program known as Kurzweil™ 3000. For each test administration, the test development contractor selects, with MSDE's approval, one test form for each content to be produced in an electronic format for verbatim reading using the Kurzweil™ software.

Limited English Proficient Students (LEPs) *(See "English Language Learners.")*

Local Accountability Coordinator (LAC) The staff member at the school district level who implements the state required assessments according to mandated procedures.

Lowest Obtainable Scale Score (LOSS) As the name states, this score is the lowest score possible. For the MSAs with a scale ranging from 240 to 650, the LOSS is 240.

Maryland Assessment Research Center for Education Success (MARCES) The Maryland Assessment Research Center for Education Success provides support to the range of assessment activities in the State, the region and nation by conducting basic and applied research to enhance the quality of assessment practice and knowledge. MARCES is a project of the Department of Measurement, Statistics, and Evaluation in the College of Education at the University of Maryland.

Maryland Report Card A collection of data compiled annually to provide information on school performance to all educational stakeholders, to support school improvement efforts, and to provide accountability at the state, local school system and school level. These data are accessible on the web at http://mdreportcard.org/.

Maryland School Assessment (MSA) The MSAs assess the Maryland content standards in mathematics, reading, and science. The reading and mathematics tests are administered annually to students in grades 3 through 8. The science test is administered annually in grades 5 and 8. MSA scores show how well Maryland students have learned the reading, mathematics, and science skills specified in the Voluntary State Curriculum (VSC). Each child receives a score in each content area that categorizes his/her performance as basic, proficient, or advanced (see Performance Level Standards). The MSAs include both selected response (multiple-choice) and brief constructed response items. All MSAs are peer reviewed by the US Department of Education and must satisfy the rigorous requirements of the federal No Child Left Behind Act.

Math Strands Seven strands are measured on the MSA math: Algebra, Geometry, Measurement, Statistics, Probability, Numbers and Computation, and Process.

Measurement Error *(See "Standard Errors of Measurement.")*

Modes of Test Administration Standardized modes for the MSAs and HSAs include the paper-and-pencil format and the computer administered format. Research should confirm that the scores resulting from these modes are comparable. For the scores to be *statistically comparable*, the scores, performance level classifications, and proficiency rates yielded by the two testing modes must be interchangeable at all levels—student, school, school district and disaggregated subgroups.

Mod-HSA An alternate assessment with modified achievement standards (AA-MAS) in reading, mathematics, and science for grades 5, 8, and 10–12 taken by the small group of students with disabilities who can make significant progress, but who may not reach grade-level achievement in the time frame covered by their IEPs.

Modification A change in the way an assessment is given or taken that alters what is being measured (reading the reading passage to a student when reading is the trait being measured). This change invalidates that student's reading score. This definition of modification is not to be confused with changes to the design and format of assessments for the 2% of students who cannot take the regular assessments with accommodations-assessments with modified academic achievement standards (MAS).

National Assessment of Educational Progress (NAEP) NAEP is the only nationally representative and continuing assessment of what America's students at grades 4, 8, and 12 know and can do in mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history. Assessments in world history and in foreign language are anticipated in 2012. There are two NAEP websites: one for program and technical information, http://nces.ed.gov/nationsreportcard, and the other for the Nation's Report Card results, http://nationsreportcard.gov.

National Psychometric Council (NPC) The name given to Maryland's technical advisory committee (TAC) that advises MSDE on complex and technical issues that relate to the continued implementation of the Maryland school Assessment Program. They also approve work done by the testing contractors including statistical analyses, equating, etc.

National Technical Advisory Council (NTAC) Sixteen members appointed by The U.S. Department of Education (USED) to advise ED on complex and technical issues that relate to the continued implementation of the No Child Left Behind Act. The chosen members come from a range of backgrounds in the assessment and accountability field, including past-and-present state education agency officials.

NCLB The No Child Left Behind Act of 2001 reauthorized the Elementary and Secondary Education Act (ESEA)—the main federal law affecting education from kindergarten through high school. Proposed by President Bush shortly after his inauguration, NCLB was signed into law on January 8th, 2002. NCLB is built on four principles: accountability for results, more choices for parents, greater local control and flexibility, and an emphasis on doing what works based on scientific research (http://www.ed.gov/policy/elsec/leg/esea02/index.html).

Non-disclosure Agreement A form signed by all education stakeholders who have access to secure test materials that affirms that they will neither discuss nor disclose to unauthorized persons the contents of the materials reviewed . Violators are subject to sanctions.

Norm-Referenced Test (NRT) Interpretation A score interpretation based on the comparison of a student's score to the scores of other students in a specific referenced population (norm group). Scores are typically reported as percentiles—the score on a test below which a given percentage of scores fall. Technically an assessment can yield several types of scores, those used for CRT and NRT interpretations, provided this goal is part of the test design. Thus, it is not the test itself that is an NRT or a CRT, per se, but the interpretation attached to the results. The design and intended use of the scores impacts the interpretation that may be attributed to them.

Operational Test Forms Actual test forms that are administered to inform an interpretation or decision that is based on test scores. If the operational test form also contains field test items embedded among the operational items, only the operational test items count towards the student's score.

Pattern Scoring Pattern scoring is a complex scoring procedure using item response theory; the weight that the item has is related to the amount of information it provides.

Public Release Test Forms MSDE selects annually one intact operational test form per content area to release. The items and student responses to CR items are posted to MSDE's mdk12.org school improvement website.

Predictive Validity The empirical degree to which test scores predict (or are related to/or correlate with) with other measures of the same content that are measured at some future time. An example is the degree to which scores on the College Board's SAT test, taken prior to entrance into college, correlates with freshman year grade point averages.

Performance Level Descriptors (PLDs) Descriptive statements for each proficiency level (Advanced, Proficient or Basic) for all content areas and grades assessed by the MSAs and HSAs that define what students know and are able to do at that level. PLDs are available on the web.

Performance Level Standards Also known as academic achievement standards, these standards are measures of performance against which yearly results are compared. Maryland has three levels of achievement:

- **Advanced** is a highly challenging and exemplary level of achievement indicating outstanding accomplishment in meeting the needs of students.
- **Proficient** is a realistic and rigorous level of achievement indicating proficiency in meeting the needs of students.
- **Basic** is a level of achievement indicating that more work is needed to attain proficiency in meeting the needs of students.

Standards help to examine critical aspects of instructional programs; help to ensure that all students receive quality instruction; hold educators accountable for quality instruction; and help to guide efforts toward school improvement. Maryland's standards were determined through deliberative processes by educators with involvement of critical stakeholders such as the legislators and members of the business community. The State Board of Education adopted all standards.

Qualifying Sets Set of student responses to a constructed response item that must be scored accurately by readers in order for them to be hired to score Maryland's test.

Range Finding The process of selecting student responses to constructed response items in order to demonstrate a range for each score level of the rubric used for scoring. Responses are judged against the anchor responses from previous administrations.

Raw Score Also known as the number correct score, the raw sore is based on the number of items answered correctly.

Reading Strands Three strands are measured on the MSA reading: General, Literary, and the Informational Reading Process.

Reliability The empirical degree to which test scores for a group of examinees are consistent or stable over repeated applications of the assessments. Reliability of test scores is a necessary, but not sufficient condition, for the validity of test score interpretations. The type of reliability typically reported for large-scale state assessments is known as internal consistency reliability. Other reliabilities, alternate-forms, test- retest, and generalizability coefficients, are not equivalent because each has a unique definition of measurement error. Estimates of reliabilities and standard errors of measurement are related—the higher the reliability, the lower the error.

Responsibilities of Test Takers Test takers are responsible for preparing for the test, following the test administration directions, representing themselves honestly on the test, and informing authorities if scores may not reflect them appropriately. See http://www.apa.org/science/ttrr.html.

Review Committees Groups of educational stakeholders who meet to perform a task in the test development process e.g., review test items for bias, determine whether test items measure the indictors they are intended to measure, score student responses to constructed response items, or recommend achievement/performance standards. The state must monitor and document the process of selecting, training, qualifying, and providing feedback to reviewers.

Rights of Test Takers Test takers have the right to receive in advance of the testing information about the nature of the tests, the intended use of the scores, and the confidentiality of the results. See http://www.apa.org/science/ttrr.html.

Rubrics A generic scoring tool that is used by trained raters to judgmentally score constructed response items (essays) according to specific criteria. The criteria include samples of student work (exemplars) at each score point.

Safe Harbor Safe Harbor allows a school that does not meet the annual performance targets for each subgroup to make AYP if the school meets all performance targets in the aggregate, and the subgroup meets the other academic indicator; and the percentage of students achieving below the proficient level in that subgroup decreases by ten percent. Safe Harbor is calculated using the last two years of test administration data.

Scale Score The scale scores for MSA reading, mathematics, and science range from 240 to 650 with a mean (average) of 400 and a standard deviation of 40. The scale score is the measurement of a student's performance relative to the three achievement levels identified by the state: Basic, Proficient, and Advanced. If a student sores 450 on the test and this score is/at or above the cut for the Proficient level, the student's performance is reported as "Proficient." Cut scores for the performance levels vary by content and grade and are shown on www.MDReportCard.org.

Scoring The method by which a numerical score is assigned to a student's response. When the test is composed of all selected response items, the student's responses recorded on a answer sheet are most frequently machined scored by electronically scanning the answer sheet. When the test is also composed of constructed response items, these items are judgmentally scored by trained readers who assign scores based of rubric criteria and a scoring guide.

Scoring Sets Multiple sets of scored student responses to constructed response items are used to train raters and monitor rater reliability during the scoring process.

- **Anchor Set** Scored papers that are clear examples of each score point on the rubric.
- **Decision Set** Scored papers that illustrate the various kinds of problems that might arise with a prompt or item.
- **Training Sets** At least two sets of up to 20 papers each, that contain a range of responses including clean papers (unambiguous examples of solid scores), line papers (papers that are on the line between one score and the next, having features of both score points), and problem papers (those with unusual approaches to the prompt that may require expert scoring). The responses are in random order of quality and unmarked.
- **Qualifying Sets** At least three sets of typically 10 responses per set that consist mostly of clean papers.
- **Calibration Sets (Validity Sets)** Sets of pre-scored student responses of mixed quality, arranged in random order, that are circulated without identification to the scorer during live scoring to establish that there is no drift in the application of scoring criteria.

Scoring Drift The failure of readers scoring constructed response items to consistently assign accurate scores to student responses.

Scoring Guides The scoring tools (rubrics or item-specific) that are used by trained raters to judgmentally score constructed response items (BCRs and ECRs) according to specific criteria defined for each score point. The scoring tools are accompanied by samples of student work (exemplars) that represent solid papers at each score point.

Section 504 Students Students who have a physical or mental impairment that substantially limits one or more major life activities, have a record of such an impairment, or are regarded as having such an impairment.

"Special Placement" Schools (LEA 24) Private institutions, state-operated programs, or non-public educational centers serving students who have been placed by the Maryland public schools. Publicly-funded students in these special placement schools *must* be included in the testing and state accountability system.

Standard Errors of Measurement (SEM) The SEM is an empirical estimate of the amount of variation that can be expected in obtained test scores for the same student if the student were to retake the test with no change in knowledge between administrations. The interpretation of the SEM is usually made in terms of a statement of probability that the score obtained by a student is within a certain distance of his or her true score (that is, the score he or she would obtain on a perfectly reliable test). The probability is .68 that a student's score will be within *one* SEM of his or her true score and .95 that it will be within *two* SEMs (assuming a normal distribution of test scores).

Standard Operating Procedures Secure and non-secure written documentation that clearly, completely, and accurately describes all facets of the test development process for multiple user groups including measurement experts, teachers, students, parents, test administrators, researchers, testing and scoring contractors, and other educational and business stakeholders. Major secure documents include, but are not limited to, detailed item and test blueprints, the actual test forms, and the test item banks. Non-secure information includes, but are not limited to, the learning standards (e.g., the core learning goals and the voluntary state curriculum), sample test item formats, test forms released post administration, annual technical reports, research studies, test administration manuals, scoring tools, score interpretation guides, and score reports.

Standards for Educational and Psychological Testing The national test standards address the professional and technical issues of test development used in education, psychology and employment. Tests must be valid, reliable and fair**. Valid tests** measure the knowledge and skills that it is supposed to measure. **Reliable tests** produce consistent scores among groups of test takers. **Fair tests** are composed of test questions that are free of bias that would disadvantage any group of test takers because of their race, ethnicity, economic status, or other characteristic. The standards are sponsored by the American Educational Research Association (AREA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) and are embraced by numerous government agencies, test publishers/ developers, credentialing boards, and academic institutions. See http://www.apa.org/science/ standards.html

Standardized Tests A standardized test is composed of the same or similar tasks or questions that are administered under the same conditions to all test takers (with the exception of those with disabilities or limited English proficiency) and scored using the same methods by the same criteria.

State Test Administration and Security Committee (STASC) The committee composed of a representative from the Attorney General's Office and measurement experts from MSDE that reviews testing irregularities that are in violation of the Code of Maryland Regulations (COMAR) Section 13A.03.04.05A: Test Administration and Data Reporting Policies and Procedures, Testing Behavior Violations and recommends sanctions such as letters of warning or letters of reprimand.

Students with Disabilities (SWDs) Students who are eligible for special education and who have current Individualized Education Programs (IEPs).

Technical Reports A technical report for each content area measured is produced annually and provides accurate, complete, current, and clear documentation of the test development process, data analyses and results for test users. Reports since 2003 for the math and reading tests, and 2007 for the science field test are available at www.marylandpublicschools.org

Test Accommodations Testing accommodations for students with disabilities (i.e., students having an Individualized Education Program or a Section 504 Plan) and students for English Language Learners (ELL) are approved and documented according to the procedures and requirements outlined in the document entitled *"Maryland Accommodations Manual: A Guide to Selecting, Administrating, and Evaluating the Use of Accommodations for Instruction and Assessment," (MAM).*

Test Administration and Coordination Manual (TACM) A detailed procedures manual for district level local accountability coordinators and school test coordinators that contains the guidelines for planning and managing a standardized administration of the assessments. Topics include test administration schedules, types of accommodations, administration monitoring, reporting test irregularities, with instructions of all permitted and prohibited activities prior to, during, and after the administration of both the computer-assisted or paper-and-pencil tests. Failure to follow these testing procedures is a violation of the Code of Maryland Regulations (COMAR) Section 13A.03.04.05A: Test Administration and Data Reporting Policies and Procedures, Testing Behavior Violations.

Test Administration Modes Maryland administers the MSAs via the traditional paper-and-pencil format and the computer-administered, on line format. For the two modes to be comparable, the scores, performance level classifications, and proficiency rates yielded by the two testing modes must be interchangeable at all levels—student, school, school district and disaggregated subgroups.

Test Bias Bias is the presence of some characteristic of an assessment item that results in the differential performance of two individuals of the same ability, but from different subgroups or with different backgrounds. Bias may be detected by judgmental or statistical procedures (e.g., DIF analyses).

Test Development Process (Summary) Typical steps in the process of high quality test development include: planning, consulting with multiple contractors, program management, test item development and review, publishing, scoring, administration, security monitoring, research, policy development, evaluation, standard setting, and communications with all stakeholders.

Test Directions Test directions specify time allowances, the nature of the test takers expected responses, and the rules regarding the use of calculators, manipulatives, dictionaries and other materials.

Test Disclosure The process of providing examples of a test form for public review. One administered test form per content area of the HSAs is released annually to the public. The sample tests are available on the web at the link http://hsaexam.org/support/practice.html

Test Equating In item response theory, *equating* is the process of equating the units and origins of two scales on which the abilities of students have been estimated from results on different tests. The process is analogous to equating degrees Fahrenheit with degrees Celsius by converting measurements from one scale to the other. The determination of comparable scores is a by-product of equating test results from equating the scales obtained from test results.

Test Examiner's Manual (TEM) A detailed procedures manual for test administrators and monitors that contains all information required for a standardized administration of the assessments. Topics include test administration schedules, types of accommodations, administration monitoring, reporting test irregularities, with instructions of all permitted and prohibited activities prior to, during, and after the administration of both the computer-assisted or paper-and-pencil tests. Failure to follow these testing procedures is a violation of the Code of Maryland Regulations (COMAR) Section 13A.03.04.05A: Test Administration and Data Reporting Policies and Procedures, Testing Behavior Violations.

Test Forms Several test forms are developed annually for each content area and grade measured. The intent of the test development process is to have forms be parallel in terms of the number of items by item format and stimulus materials, and the balance of the content covered. All forms are reviewed and approved by the contractor and MSDE assessment staff. One form for each content and grade is the designated Braille/large print form.

Test Item Construction Flaws Test item characteristics that limit an item's ability to measure the intended outcome. Examples include, more than one defensible correct answer when only one answer may be selected, no defensible correct answer, one or more implausible nonfunctioning distracters, clues to the correct answer within an item or among items, and confusing or ambiguous wording that prevents knowledgeable test takers from responding correctly. Item analysis data and the expert judgments of content and measurement specialists detect these flaws before the items are used on operational tests.

Testing Incident Report Form The form that local accountability coordinators complete and submit whenever a test security violation or test administration procedural deviation (Category 2 violation) from MSDE testing policy occurs.

Test Resources Test information is available in multiple foreign translations at http://www.marylandpublicschools.org/MSDE/newsroom/publications/pubsother/.

Test Security Access to specific test content is limited to only those who have a need to know it for test development, test scoring, and test evaluation. Stakeholders who have access to secure test materials (e.g., test items and test forms) sign non-disclosure agreements. Examinees or educators who intentionally disclose test content are subject to having their test scores invalidated or other sanctions identified by the State Test Administration and Security Committee.

Test Specifications A detailed description for the design and development of a test that specifies the number and/or proportion of items by item format that should measure each content and skill area, and the desired measurement statistics (e.g., item difficulties and discrimination) of the items and the test. This is typically a secure document.

Test-wiseness Test-taking strategies students use to perform optimally on a test. These include, but are not limited to, the following: read and follow test directions carefully, read each question and all the answer choices carefully before choosing the best answer, mark answers as directed, use time efficiently, answering easy questions first, take risks and guess if there's no penalty for incorrect answers, and change answers if they might have been answered incorrectly.

The 1602 Combined Passing Score Maryland permits students to meet the HSA graduation requirement by either passing all four tests or by achieving 1602 total score points across the four tests—a number representing the combined passing scores for the four tests. (A student's cumulative score on the tests in algebra, biology, government, and English 2 must equal or exceed 1602 points).

Understanding By Design An educational taxonomy with six facets of understanding used for the design of instruction and assessments. According to Wiggins and McTighe, in *Understanding by Design, o*ne truly understands when one **Can Explain:** provide thorough, supportable and justifiable accounts of phenomena, facts and data; **Can Interpret:** tell meaningful stories; offer apt translations; provide a revealing historical or personal dimension to ideas and events; **Can**

Apply: effectively use and adapt what we know in diverse contexts; **Have perspective:** see and hear points of view through critical eyes and ears; see the big picture; **Can empathize:** find value in what others might find odd, alien, or implausible; perceive sensitively on the basis of prior direct experience; **Have self-knowledge:** perceive the personal style, prejudices, projections, and habits of mind that both shape and impede our own understanding; aware of what we do not understand and why understanding is so hard (http://www.greece.k12.ny.us/instruction/ela/6-12/BackwardDesign/BDstep1.htm).

Validity The degree to which accumulated evidence and theory support specific interpretations of the test scores proposed by the test developers. Validity evidence can be extensive and diverse. It may be based on the test's content, test taker response processes, the test's internal structure, the test's relationship to external variables, and the intended and unintended consequences of testing. Traditional nomenclature for validity includes the terms: concurrent validity, consequential validity, construct validity, content validity, convergent validity, criterion-related validity, discriminant validity, face validity, and predictive validity.

Validity Sets Sets of student responses that are circulated during live scoring to establish that there is no drift (change over time from day 1 to the final day of scoring) in the application of scoring criteria. Validity materials are pre-scored and circulated without identification to the scorer.

Voluntary State Curriculum (VSC) The document that identifies the Maryland Content Standards and aligns them with the Maryland Assessment Program. The curriculum documents are formatted so that each begins with content standards or broad, measurable statements about what students should know and be able to do. The next level of specificity, indicator statements, narrow the focus for teachers. Finally, the objectives provide teachers with very clear information about what specific learning should occur. When objectives are measured on the MSA in Mathematics and Reading, each is followed by an assessment limit. For details by subject and grade: http://www.mdk12.org/assessments/vsc/index.html

Websites For additional information, visit www.MdReportCard.org, www.HSAexam.org, and www.MdK12.org**.**

Appendix B

Acronyms

AA-AAS	Alternate Assessment based on *Alternate* Academic Achievement Standards
AA-MAS	Alternate Assessment based on *Modified* Academic Achievement Standards
AAS	Academic Achievement Standards
AERA	American Educational Research Association
Alt-MSA	Alternate Maryland School Assessment *(See AA-AAS.)*
AMO	Annual Measurable Objective
APA	American Psychological Association
AYP	Adequate Yearly Progress
BOE	Board of Education
CAT	Computer Adaptive Tests or Computer Administered Test
CCSSO	Council of Chief State School Officers
CLGs	Core Learning Goals
COMAR	Code of Maryland Regulations Section 13A.03.04.05A: Test Administration and Data Reporting Policies and Procedures, Testing Behavior Violations.
CR	Constructed Response Item
CRT	Criterion-referenced Test Interpretation
DAA	Division of Assessment and Accountability, MSDE
DIF	Differential Item Functioning
DIS	Discrimination Index

ELLs	English Language Learners
FAQs	Frequently Asked Questions
FT	Field Test
HOSS	Highest Obtainable Scale Score
HSA	High School Assessment
ICC	Item Characteristic Curve
IRT	Item Response Theory
LEA	Local Education Agency
LEP	Limited English Proficient students
LAC	Local Accountability Coordinator
LOSS	Lowest Obtainable Scale Score
LSS	Local School System
MAM	Maryland Accommodations Manual: A Guide to Selecting, Administrating, and Evaluating the Use of Accommodations for Instruction and Assessment.
MARCES	Maryland Assessment Research Center for Education Success, at the University of Maryland
MC	Multiple-choice item
MFTP	Maryland Functional Testing Program
Mod-HSA	Modified High School Assessment *(See AA-MAAS.)*
MSA	Maryland School Assessment
MSDE	Maryland State Department of Education
MSPAP	Maryland School Performance Assessment Program

NAEP	National Assessment of Educational Progress
NCLB	No Child Left Behind Act of 2001
NCME	National Council on Measurement in Education
NPC	National Psychometric Council
NRT	Norm-referenced Test Interpretation
PLD	Performance Level Descriptor
p-value	Item difficulty
R-bis	Point bi-serial correlation
SEM	Standard Error of Measurement
SPR	Student-produced Response
SR	Selected Response item
TAC	Technical Advisory Committee
TACM	Test Administration and Coordination Manual
TEM	Test Examiner's Manual
TIRF	Testing Incident Report Form
USED	U.S. Department of Education
VSC	Voluntary State Curriculum

Appendix C: Quick Reference Guide to Accommodations for Students with Disabilities (SWD)

NOTE: Users of this Appendix must have the complete text of the 2007-2008 Maryland Accommodations Manual available for reference.

1. SWD Presentation Accommodations

Visual Presentation Accommodations	Conditions for Use In Instruction and Assessment
I-A: Large Print	I , A
I-B: Magnification Devices	I , A
I-C: Interpretation/Transliteration for the Deaf or Hard of Hearing	I , A
Tactile Presentation Accommodations	
I-D: Braille	I , A
I-E: Tactile Graphics NOTE: For purposes of State assessments, any tactile graphics needed are included with the Braille version of the test.	I , A*
Auditory Presentation Accommodations	
I-F: Human Reader, Audio Tape, or Compact Disk Recording for Verbatim Reading of Entire Test	I , A*
I-G: Human Reader, Audio Tape, or Compact Disk Recording for Verbatim Reading of Selected Sections of Test	I, A*

* Use of the verbatim Reading accommodation is permitted on all assessments as a standard accommodation, with the exception of:

(1) the Maryland School Assessment (MSA) in Reading , grade 3 ONLY, which assesses a student's ability to decode printed language. Students in grade 3 receiving this accommodation on the assessment will receive a score based on standards 2 and 3 (comprehension of informational and literary Reading material) but will not receive a score for standard 1, general Reading processes; and

(2) the Maryland Functional Reading Test.

1. SWD Presentation Accommodations (continued)

Auditory Presentation Accommodations *(continued)*	Conditions for Use In Instruction and Assessment
1-H: Audio Amplification Devices	I , A
1-J: Books on Tape	I , N/A
1-K: Recorded Books	I , N/A
Multi-Sensory Presentation Accommodations	
1-L: Video Tape and Descriptive Video NOTE: No Maryland assessments currently incorporate video-taped stimulus materials. However, if video tape is used, students must have access to dosed captioning on video materials, as appropriate.	I , A
1-M: Screen Reader for Verbatim Reading of Entire Test	I , A*
1-N: Screen Reader for Verbatim Reading of Selected Sections of Test	I , A*
1-0: Visual Cues	I , A
1-P: Notes, Outlines, and Assessments	I , N/A
1-0: Talking Materials	I , A
Other Presentation Accommodations	
1-R: Other	Determined on a case-by-case basis in consultation with MSDE

* Use of the verbatim Reading accommodation is permitted on all assessments as a standard accommodation, with the exception of:
(1) the Maryland School Assessment (MSA) in Reading, grade 3 ONLY, which assesses a student's ability to decode printed language. Students in grade 3 receiving this accommodation on the assessment will receive a score based on standards 2 and 3 (comprehension of informational and literary Reading material) but will not receive a score for standard 1, general Reading processes; and
(2) the Maryland Functional Reading Test.

Any screen reader may be used for instruction, but the only screen reader currently supported by the State for assessment is the Kurzweil™ 3000. In order for students to use the Kurzweil™ 3000 screen reader for testing, students must have used a screen reader in instruction and have had an opportunity to become familiar with the operation of the Kurzweil™ 3000 interface. Although a Human reader is always permissible to deliver a verbatim Reading accommodation, the State encourages the use of screen readers on State testing, to promote standardization of the verbatim Reading accommodation.

2. SWD Response Accommodations

Response Accommodations	Conditions for Use In Instruction and Assessment
2-A: Scribe	I , A
2-B: Speech-to-Text	I , A
2-C: Large-Print Response Book	I , A
2-D: Brailler	I , A
2-E: Electronic Note-Takers and Word Processors	I , A
2-F: Tape Recorder	I , A
2-G: Respond on Test Booklet	I , A
2·H: Monitor Test Response	I , A
Materials or Devices Used to Solve or Organize Responses	
2-J: Calculation Devices	I , A
2-K: Spelling and Grammar Devices	I , A*
2-L: Visual Organizers	I , A**
2-M: Graphic Organizers	I , A
2-N: Bilingual Dictionaries	I , A
Response Accommodations	
2-0: Other	Determined on a case-by-case basis in consultation with MSDE

* Spelling and grammar devices are not permitted to be used on the English High School Assessment.

** Photocopying of secure test materials requires approval and must be done under the supervision of the LAC. Photocopied materials must be securely destroyed under the supervision of the LAC. Use of highlighters may be limited on certain machine-scored test forms, as highlighting may obscure test responses. Check with the Test Administration and Coordination Manual (TACM) for each test or consult with the LAC before allowing the use of highlighters on any State test.

3. SWD Timing and Scheduling Accommodations

Timing and Scheduling Accommodations	Conditions for Use In Instruction and Assessment
3-A: Extended Time	I , A
3-B: Multiple or Frequent Breaks	I , A
3-C: Change Schedule or Order of Activities – Extend Over Multiple Days	I , A
3-D: Change Schedule or Order of Activities – Within One Day	I , A
Other Timing and Scheduling Accommodations	
3-E: Other	Determined on a case-by-case basis in consultation with MSDE

4. SWD Setting Accommodations

Setting Accommodations	
4-A: Reduce Distractions to the Student	I , A
4-B: Reduce Distractions to Other Students	I , A
4-C: Change Location to Increase Physical Access or to Use Special Equipment – Within School Building	I , A
4-D: Change Location to Increase Physical Access or to Use Special Equipment – Outside School Building	I , A
Setting Accommodations	
4-E: Other	Determined on a case-by-case basis in consultation with MSDE

Appendix D: Quick Reference Guide to Accommodations for English Language Learners (ELLs)

1. ELL Presentation Accommodations

Auditory Presentation Accommodations	Conditions for Use In Instruction and Assessment
1-F: Human Reader, Audio Tape, or Compact Disk Recording for Verbatim Reading of Entire Test	I , A*
1-G: Human Reader, Audio Tape, or Compact Disk Recording for Verbatim Reading of Selected Sections of Test	I , A*
Multi-Sensory Presentation Accommodations	
1-K: Recorded Books	I , N/A
1-L: Video Tape and Descriptive Video	I , N/A
1-M: Screen Reader for Verbatim Reading of Entire Test	I , A**
1-N: Screen Reader for Verbatim Reading of Selected Sections of Test	I , A**
1-P: Notes, Outlines, and Assessments	I , N/A
1-Q: Talking Materials	I , A
Other Presentation Accommodations	
1-R: Other	Determined on a case-by-case basis in consultation with MSDE

* Use of the verbatim Reading accommodation is permitted on all assessments as a standard accommodation, with the exception of
(1) the Maryland School Assessment (MSA) in Reading, grade 3 ONLY, which assesses a student's ability to decode printed language. Students in grade 3 receiving this

accommodation on the assessment will receive a score based on standards 2 and 3 (comprehension of informational and literary Reading material) but will not receive a score for standard 1, general Reading processes;
(2) the Maryland Functional Reading Test and
(3) the ELPT. This accommodation is not permitted for students classified as RELL unless that student is also SWD, in which case the accommodation may be allowable, based on the student's IEP.

** This accommodation is not permitted for a student identified as RELL unless that student is also SWD, in which case the accommodation may be allowable, based on the student's IEP. Use of the verbatim Reading accommodation is permitted on all assessments as a standard accommodation, with the exception of
(1) the Maryland School Assessment (MSA) in Reading, grade 3 ONLY, which assesses a student's ability to decode printed language. Students in grade 3 receiving this accommodation on the assessment will receive a score based on standards 2 and 3 (comprehension of informational and literary Reading material) but will not receive a score for standard 1, general Reading processes, and
(2) the Maryland Functional Reading Test.

Any screen reader may be used for instruction, but the only screen reader currently supported by the State for assessment is the Kurzweil™ 3000. In order for students to use the Kurzweil™ 3000 screen reader for testing, students must have used a screen reader in instruction and have had an opportunity to become familiar with the operation of the Kurzweil™ 3000 interface. Although a human reader is always permissible to deliver a verbatim Reading accommodation, the State encourages the use of screen readers on state testing, to promote standardization of the verbatim Reading accommodation.

2. ELL Response Accommodations

Response Accommodations	Conditions for Use In Instruction and Assessment
2-A: Scribe	I , A
2-E: Electronic Note-Takers and Word Processors	I , A
2-F: Tape Recorder	I , A
2-G: Respond on Test Booklet	I , A
2-H: Monitor Test Response	I , A
Organize Responses	
2-K: Spelling and Grammar Devices	I , A*
2-N: Bilingual Dictionaries	I , A
Other Response Accommodations	
2-Q: Other	Determined on a case-by-case basis in consultation with MSDE

* Spelling and grammar devices are not permitted to be used on the English High School Assessment.

3. ELL Timing and Scheduling Accommodations

Timing and Scheduling Accommodations	Conditions for Use In Instruction and Assessment
3-A: Extended Time	I , A
3-B: Multiple or Frequent Breaks	I , A
3-C: Change Schedule or Order of Activities – Extend Over Multiple Days	I , A
3-D: Change Schedule or Order of Activities – Within One Day	I , A
Other Timing and Scheduling Accommodations	
3-E: Other	Determined on a case-by-case basis in consultation with MSDE

4. ELL Setting Accommodations

Setting Accommodations	
4-A: Reduce Distractions to the Student	I , A
4-B: Reduce Distractions to Other Students	I , A
4-C: Change Location to Increase Physical Access or to Use Special Equipment – Within School Building	I , A
Other Setting Accommodations	
4-E: Other	Determined on a case-by-case basis in consultation with MSDE