Maryland High School Assessment

May 2001 Administration

Scoring Report

Measurement Incorporated
423 Morris St
Durham NC 27701

October 2001

Introduction

The Maryland High School Assessment Field Test, May 2001 administration, consisted of multiple forms with multiple constructed response items for each of the five content areas assessed: Algebra/Data Analysis, Biology, English, Geometry, and Government. Measurement Incorporated (MI) scored constructed response items in each form and content area of the High School Assessment using Maryland's content specific generic rubrics and scoring guides and training sets. The guides and sets consisted of responses selected by Maryland educators to define acceptable limits within each score point descriptor range and were compiled by MI Scoring Directors for each specific content area and test. A guide and practice sets were created for each item in the assessment with annotations to link the rubric to the specifics of the student response, thus providing the rationale for the score.

Each test was then scored by professional Scorers who were systematically trained by each Scoring Director using the above-mentioned materials. Scores were recorded on customized score sheets. The score sheets were scanned at the scoring sites using Opscan optical scanners. Scores were transmitted electronically to MI's Information Technology Department, merged into data files, and sent to CTB/McGraw-Hill, the development contractor. Across all content areas, forms, and items, 212,028 answer books were hand-scored. Data was collected from approximately 333,645 individual score sheets.

As an adjunct to the statistical evaluation of items, Scorers used an item evaluation sheet to record trends and idiosyncrasies observed when scoring student papers. Each Scoring Director discussed each item with the Scorers, read the evaluation sheets, added his/her own observations, and compiled an anecdotal report on item scoring. This additional item evaluation information was supplied to MSDE Content Specialists.

In every aspect of the May 2001 MDHSA scoring conducted by MI, there was a collaborative effort with key staff in the MSDE. The Scoring Lead and Content Specialists were consulted for all decisions, creating the scoring process that Maryland directed and that MI executed.

Rangefinding

Constructed responses are an integral part of Maryland's High School Assessment Program. Maryland is strongly invested in a program that will link the HSA to the performance assessment of its MSPAP assessment program (grades 3,5,8). The test designers were given the charge to include constructed response items in the assessment to provide student-produced evidence of application and reasoning as valued in Maryland's educational programs and strategies.

To guide the scoring of constructed responses—and more generally to provide a visible performance goal to students, teachers and Scorers—committees of Maryland educators constructed content specific generic rubrics. Upon the administration of each new test

item, the generic rubric becomes item specific through a process that will be referred to here as rangefinding. This process is pivotal in the success of Maryland's testing program. It calls upon the expertise of Maryland educators in concert with the scoring contractor's professional staff. It is the foundation of constructed response scoring. Committees composed of Maryland educators, along with leadership personnel from the contractor's staff, met prior to the May 2001 scoring of constructed responses to pre-score hijacked responses from the current administration. The committees were content specific: English, Algebra/Data Analysis, Geometry, Government, and Biology. By first training on generic rubrics and established "anchors," or samples from previous administrations, the committee calibrated their scores of student responses to scores from previous administrations. Committees then proceeded to score each new item in the field test using the generic rubrics and anchor papers.

The Maryland educators produced scored responses for each item that would become the referenced criteria for the rest of the scoring for those items. Academic discussions of the criteria and the student responses led to a consensus of scores for each score level on the rubrics. Scoring guides and training sets made up of committee-scored papers became the blueprint of the scoring process. All scores assigned throughout the process were based on the foundation laid by these committees. The statistics generated from this scoring would determine if an item should be entered into the HSA item pool for future operational test administrations.

After each committee was trained and calibrated on previously scored HSA test items, the committees were then broken down into sub-groups. English, Algebra/Data Analysis, and Geometry separated into a committee for Extended Constructed Responses (ECRs) and one for Brief Constructed Responses (BCRs). Government separated into a committee for ECRs and two committees for BCRs.

Biology has only BCRs but still separated into two small committees because of the large number of items to be scored. One of the two smaller committees used an item-specific process to determine scores, diverging from the anchor-based scoring of the other committee. Results were not parallel between these two diverse methods, so Biology conducted an extra rangefinding session to review some items and to reassign scores based on anchor papers. This was a very time consuming, yet productive, process with many benefits. A true item evaluation revealed strong and weak items, and a closer look at the constraint on item writing to produce items at the indicator level revealed a restriction counterproductive to constructive response item richness. The result was an opening up of item writing to the goal level for Goal One and the expectation level for Goal Three. Biology items can now more realistically measure a student's ability to use scientific reasoning to make connections and assessments using biological principles.

Rangefinding Activities

Government
May 29 – June 6 (BCR); May 30- June 6 (ECR)
3 Forms, 27 BCR items, 3 ECR items

English
May 31- June 5 (BCR); May 31 – June 6 (ECR)
5 Forms, 14 BCR items, 5 ECR items

Biology
June 4 – June 9
4 Forms, 34 BCR items

Second Biology Rangefinding Session

July 18 – July 20
3 Forms, 11 BCR items

Geometry
June 6 – June 9
3 forms, 14 BCR items, 12 ECR items

Algebra/Data Analysis
June 11 – June 15
5 Forms, 19 BCR items, 9 ECR items

Preparation for Rangefinding

The same day that each test was administered, a sample shipment of each test from schools randomly selected by MSDE was express shipped to the Measurement Incorporated Central Office. This sample is referred to as a "hijack." The responses in these tests were carefully reviewed by MI specialists, in accordance with the generic rubrics, to select a variety of responses for Maryland educators to evaluate. The selected responses were assembled in packets that contained an adequate number of responses to show the full range of the hijack sample and a variety of student approaches to each test item. These responses were duplicated to provide a copy for each committee member. It is assumed that the hijack would be representative of the whole assessment. However, when a new approach to a response occurs in the actual scoring, MI always consults the MSDE Scoring Lead and MSDE Content Facilitators for direction. MI is diligent to implement Maryland decisions and has immediate access to MSDE Scoring and Content Facilitators when "new," inevitable questions occur during scoring.

All copying, printing, and shipping functions were carried out by MI, and all materials were kept secure throughout the process.

Preparation of Training Materials

Upon the completion of rangefinding, Scoring Directors used committee-scored responses to create Scoring Guides and Training Sets that were unique to each item. These were used in conjunction with the rubrics to train Team Leaders and Scorers.

One guide and two training sets were created for each item. Examples of responses at each scorepoint were included in the Scoring Guide in scorepoint order. The score assigned in rangefinding and a brief annotation was written on each response in the guide. In contrast, examples in the training sets were in random scorepoint order, with no score or annotation. These sets were given to the Scorers after they were trained on the guide. Scorers used the guide and rubric to assign scores to the training set responses.

Training and Hand Scoring

Training and hand scoring took place at four satellite locations. The number of Scorers required for each of the five content areas made it necessary to utilize multiple scoring center locations to ensure that an adequate number of Scorers and Team Leaders would be available and could be assigned to complete scoring in the time allotted.

MI Scoring Staff

Scoring Project Management

The function of MI Scoring Project Management is to coordinate and execute all hand scoring and related activities for the project. Project Management works closely with MSDE content and scoring personnel and acts as liaison between MSDE and the Content Area Scoring Directors.

Content Area Scoring Directors

Each Scoring Director participated in rangefinding, selected training papers, prepared scoring guides, trained and monitored Scorers as well as Team Leaders, annotated papers, and were responsible for all operations necessary for conducting a successful project. Additionally, each of the Scoring Directors had education and experience in the content area to which they were assigned.

Team Leaders

In selecting team leaders, MI's management staff and the scoring directors reviewed the files of all available scoring staff. They looked for people who were experienced team leaders with a record of good performance on past projects, as well as Scorers who had been recommended for promotion to team leader.

Effective Scorer training and accurate scoring relies to a great extent on having knowledgeable, flexible team leaders. Team leaders assisted in training Scorers in team discussions of training sets, and were responsible for distributing, collecting, and

accounting for training packets and sample papers during each scoring session. During scoring, team leaders responded to questions, spot-checked scores assigned by Scorers, and counseled or retrained Scorers having difficulty. Team leaders also monitored the scoring patterns of each Scorer throughout the project, conducted retraining as necessary, and helped to maintain a professional working environment.
In addition to one team leader per team of 8 to 10 Scorers, each scoring director had a floating team leader, or room coordinator. This person directly assisted the scoring director in maintaining paper flow and supervising team leaders, and assisted other team leaders in monitoring Scorer performance during training and scoring.

Scorers
Because MI has been conducting writing and performance assessment scoring for many years, we already had available a pool of qualified, experienced Scorers at our established scoring centers. MI routinely maintains supervisors' evaluations and performance data for each person who works on each scoring project in order to determine employment eligibility for future projects. We employed many of our experienced Scorers for this project and recruited new ones as well.

MI procedures for selecting new Scorers are very thorough. After advertising in local newspapers, with the job service, and elsewhere, and receiving applications, staff in our human resources department review the applications and then schedule interviews for qualified applicants. Qualified applicants are those with a BA or BS in English, language arts, education, mathematics, science, social studies, or a related field. Each qualified applicant must pass an interview by experienced MI staff, write an acceptable essay, and receive good recommendations from references. We then review all the information about each applicant and either offer employment or inform the applicant of nonacceptance.

Each scoring center has an operational supervisor (Site Manager) who recruited Scorers, oversaw the secure receipt, storage, and delivery of all scoring materials and student responses, and who supervised warehouse and clerical personnel involved in the scoring project.

The scheduling of start dates for the training of the ECR and BCR groups for each content area was staggered so that the Scoring Director for each content area was able to train each group and briefly monitor early progress before moving on to train the next group. The group of trained Scorers worked under the direction of the Room Coordinator, or Floating Team Leader, who forwarded all scoring decisions to the Scoring Director.

Team Leader and Scorer Training

The following procedures for Team Leader and Scorer training were used for all content areas at all scoring centers.

The Scoring Director for each content area conducted the training of the Team Leaders and Scorers. All Team Leaders and Scorers were trained using the rubrics approved by

the MSDE, along with anchor, or guide, papers and training papers scored by committee during the rangefinding meetings. Scorers were assigned to a scoring group consisting of one Team Leader and 8 to 12 Scorers. Each Team Leader and Scorer was assigned an individual number for easy identification of his or her scoring work throughout the scoring session.

Training was orchestrated so that Scorers understood how to apply the MSDE rubric and criteria in scoring the papers, learned how to reference the scoring guides, and developed the flexibility needed to deal with a variety of responses and to retain the consistency needed to score all papers accurately. In addition to the initial scoring training, a significant amount of time is allotted for demonstrations of paper flow, explanations of "alerts" and "flagging," and instructions about other procedures that are necessary for the conduct of a smooth project.

Since the assessment consisted of multiple forms per content area, training continued throughout the project. Items were scored in sets of three or four per form, and a separate training session was held for each new set of items to be scored.

Scoring of Pre-operational Assessment

The following procedures for scoring were used at all scoring centers:

Student responses were received at MI's Headquarters for processing. Following a security check-in scan, the individual student answer booklets were processed into packets of student responses, with machine-scanable score sheets, or scan sheets. These were sent via secure carriers to the appropriate scoring locations for each content area. Upon arrival at the scoring centers, each shipment was checked for completeness, inventoried, and securely warehoused on site.

After Scorers had been trained on a given set of items, packets of student answer documents within a form were distributed randomly to the Scorers by team. Ten percent of the packets were read twice. These packets contained two score sheets, one for each reading. Special care was taken to ensure that the packets identified for second reading were distributed equally among the entire pool of Scorers. However, no second reading packets were distributed to the same team of the Scorer who did the first reading. Also, the second Scorer used a separate score sheet and was unaware of scores assigned by the first Scorer.

As a Scorer completed a packet of papers, he or she returned it to the envelope and returned it to the Team Leader, along with the score sheet. The clerical aide picked up completed packets and score sheets from team leaders. Score sheets collected by clerical staff were visually checked for errors such as missing bubbles or extra bubbles, then sent to be scanned. The scanner was programmed to automatically reject any score sheet that was incompletely or improperly bubbled. These rejected score sheets were then matched up with the appropriate packet of responses, and returned to the Scoring Director for rescoring. Ten percent of all responses received a second reading. Aides redistributed the

packets designated for second readings. The procedure for the second reading was the same as that for the first reading, except that the second Scorer used the second score sheet in the envelope. As with the first score sheets, the second score sheets were scanned and the scores merged into the database.

Types of responses that were not anticipated and that could not be scored using the range finding examples were forwarded to the Project Director and Assistant Project Director by the Scoring Directors. After brief review, project management then forwarded these responses to MSDE scoring and content specialists for scoring decisions. These decisions and the accompanying explanations from MSDE were then given to the Scoring Directors. In this way, responses with a new and unanticipated approach to the question or otherwise aberrant responses could be scored and these examples used as scoring tools (guide papers) to score responses with similar strategies. All "decision" responses were documented for the permanent record.

Alerts were handled in a similar fashion. In training, Scorers were advised to flag responses that may indicate teacher interference, plagiarism, suicidal threats or other threats, or parental or other abuse. They submitted such responses immediately to their Team Leader or to the Scoring Director. At that point, the Scoring Director submitted a copy of the student response and an accompanying alert form to project management in Durham. Project management then requested identifying student information for the response. This information, along with the copy of the response, was then forwarded to MSDE for follow up.

The following is an overview of training and handscoring activities by content area and scoring center location:

ENGLISH

Dates of Activity: June 18 through August 10

ENGLISH: NUMBER OF FORMS, ITEMS, ANSWER BOOKLETS
May 2001

Form		BCR Items	ECR Items	Total Items/Form	N-Count*
K		3	1	4	8,824
M		3	1	4	8,629
N		2	1	3	8,458
P		3	1	4	8,322
Q		3	1	4	8,144
R		2	1	3	7,861
TOTALS	**6**	**6**	**6**	**22**	**50,238**

*N-Count is the number of student responses read, per item, plus10% second readings.

ENGLISH: NUMBER OF SCORERS
May 2001

Number of Scorers BCR	Number of Scorers ECR	Total Number of Scorers
49	23	72

English hand scoring was not completed before the deadline. This was due to the demands of the training tasks as well as a packet processing and delivery problem that originated at MI in Durham. These problems indicate to MI the need to increase staffing and improve warehouse and IT procedures for the size, scale, and structure of the May administrations of the MDHSA. Our IT Department has hired additional programming staff for the MDHSA and is also improving the quality of scanning equipment at our Durham office.

ALGEBRA/DATA ANALYSIS

Dates of Activity: June 25 through August 16

ALGEBRA: NUMBER OF FORMS, ITEMS, ANSWER BOOKLETS
May 2001

Form		BCR Items	ECR Items	Total Items/Form	N-Count
L		6	4	10	9,901
M		5	3	8	9,765
N		6	4	10	9,557
P		5	3	8	9,408
Q		7	3	10	9,204
R		7	3	10	9,052
TOTALS	**6**	**36**	**20**	**56**	**56,887**

*N-Count is the number of student responses read, per item, plus10% second readings.

ALGEBRA: NUMBER OF SCORERS
May 2001

Number of Scorers BCR	Number of Scorers ECR	Total Number of Scorers
41	39	80

Algebra/Data Analysis items generally had a large percentage of blanks, refusals, and otherwise nonscorable responses. Form L, Item 15 is the exception with only 5.1%, but other items ranged from 15% to over 50% nonscorables.

The Scoring Director notes that students did not seem to understand "justification" and "explanation." Students appeared to think that explanation, or mathematical support for their answers, would suffice for justification.

BIOLOGY

Dates of Activity: June 25 through August 29

BIOLOGY: NUMBER OF FORMS, ITEMS, ANSWER BOOKLETS
May 2001

Form		Constructed Response Items (No ECRs)	N-Count
J		9	9,637
L		7	9,428
M		9	9,168
N		9	8,883
P		9	8,653
TOTALS	**5**	**43**	**45,769**

*N-Count is the number of student responses read, per item, plus10% second readings.

BIOLOGY: NUMBER OF SCORERS

Number of Scorers	**64**

Scoring was not completed by the deadline. Late rangefinding results delayed each stage of the process. Because of the resulting extended timeline, attrition of Scorers became significant in the two necessary additional weeks of scoring. This further slowed the scoring process at a point when it was not feasible to add and train new Scorers.

GEOMETRY

Dates of Activity: June 25 through August 1

GEOMETRY: NUMBER OF FORMS, ITEMS, ANSWER BOOKLETS
May 2001

Form		BCR Items	ECR Items	Total Items/Form	N-Count
N		5	4		8,724
Q		5	4		8,497
R		5	4		8,306
S		4	4		8,137
Totals	**4**	**19**	**16**		**33,664**

*N-Count is the number of student responses read, per item, plus10% second readings.

GEOMETRY: NUMBER OF SCORERS
May 2001

Number of Scorers BCR	Number of Scorers ECR	Total Number of Scorers
27	31	58

The Scoring Director noted that, similarly to Algebra/Data Analysis, students had apparent difficulty with "justification" and "explanation."

GOVERNMENT

Dates of Activity: June 18 through August 10

GOVERNMENT: NUMBER OF FORMS, ITEMS, ANSWER BOOKLETS
May 2001

Form		BCR Items	ECR Items	n-count
K		9	1	12,754
M		9	1	12,418
N		9	1	12,208
P		9	0	11,649
Totals	**4**	**36**	**3**	**49,020**

*N-Count is the number of student responses read, per item, plus10% second readings.

GOVERNMENT: NUMBER OF SCORERS

Number of Scorers BCR	Number of Scorers ECR	Total Number of Scorers BCR and ECR
44	35	79

This content area had two BCR scoring teams in addition to a third scoring team for ECRs. The Scoring Director's task was challenging as she prepared materials, trained each group, and monitored scoring for each group. This indicates the need for additional scoring leadership for this content area.

Quality Control of Scoring

A concern regarding the scoring of any open-response test is the reliability and accuracy of the scoring. Several procedures ensured quality control on the HSA. The first of these was successful rangefinding meetings. Consistent rangefinding scoring leads to smooth Scorer training, which as a result, enhances the accuracy of scoring.

A second quality control mechanism was the experience of the leadership personnel in conducting the training and scoring sessions. MI's content area Scoring Directors were skilled at conducting initial Scorer training and qualifying procedures and were successful in schooling Scorers on how to score a variety of responses and still hold to the criteria, as well as how to handle unusual responses. Part of this process was establishing good lines of communication between Scoring Directors, Team Leaders, and Scorers.

Third, all Scoring Directors, all Team Leaders, and usually most of the Scorers at MI's current facilities have had previous experience on large-scale scoring projects. While new Scorers cannot be expected to have had prior scoring experience, all Scorers were trained to implement the scoring criteria and to maintain consistent and reliable scoring throughout the project.

Fourth, unbiased scoring was ensured because the only identifying information on the student papers is the identification number. Unless the students signed their names, wrote about their hometowns, or in some way provided other identifying information, the Scorers had no knowledge of them. The unavailability of identifying information on the papers helped to ensure unbiased scoring.

Finally, the quality of each Scorer's work was constantly monitored during the project:

Scoring Directors identified scoring trends of individual Scorers during the initial training process and had Team Leaders spot-check Scorers during scoring of "live" packets throughout the scoring process. This spot-checking was a major responsibility of Team Leaders through the entire course of the project.

Ten percent of all constructed response items received a second reading. From matching the scores to those of the first reading, valuable information could be gathered regarding Scorer agreement rates and scoring trends. Scorer status reports were generated for review by the Scoring Directors and Project Managers, who are experienced in using them to identify Scorers having difficulty, as well as to identify specific items causing problems for the whole room in general.

Spot-checking and status reports provided project management with continuous feedback not only on individual Scorers but also on room-wide scoring trends. Scoring Directors met throughout the day with Team Leaders and, using daily status reports, questions posed by Scorers, and observations from spot-checking, devised retraining strategies to keep Scorers on task with MSDE criteria.

Retraining strategies were geared to the type and degree of scoring difficulty that a Scorer may have been experiencing, and were implemented to address scoring problems on an individual basis. For example, if a Scorer displayed a pattern of scoring errors (i.e., scoring either too high or too low), the Team Leader reviewed and discussed with the Scorer the anchor papers and criteria applicable to the problematic score point line(s). If a Scorer seemed to be scoring erratically (i.e., no discernable pattern of errors), a more

intensive review of the overall criteria was required, facilitated by discussion with the Scorer to pinpoint the element(s) of the criteria that may have been causing confusion.

Team Leaders also discussed the results of Scorer status reports on an individual basis with Scorers whose performance was in need of improvement and examined the score sheets of those Scorers to ensure that adherence to the criteria was being maintained. For Scorers who were experiencing particular difficulty, the Team Leader acted as a "reading partner" for a packet or two, scoring the papers along with the Scorer in order to point out elements of the papers to look for when assigning scores, to provide a direct example of how to approach the responses, and to discuss with the Scorer the most effective ways to apply the scoring criteria. Because it is rather time-consuming, the "reading partner" strategy was generally reserved for Scorers whose scoring had still not improved sufficiently after other retraining methods had been tried. If consistent scoring still could not be achieved, the Scorer was dismissed.

Dismissals were rare for this assessment, and only five Scorers total were released from the project due to inability to score accurately: one from the Algebra group, two from the English group, and two from Biology.

In future administrations of the High School Assessment, with individual scores reported at the student level, all responses will receive a second reading to ensure the accuracy of the score of each individual student response. In the case of a two-point disagreement in scores, a third (resolution) reading will be done by an expert Scorer.

In addition, qualifying and validity sets will be included in future scoring sessions for individual accountability. The qualifying sets will be similar in design to the training sets already used. Team Leaders and Scorers will have to meet a minimum standard of performance, that is, a minimum percentage of agreement with the "true" scores of responses in the qualifying set, in order to work on the project. Validity sets containing responses with "true" scores, also similar in design to training and qualifying sets, will be administered to Scorers throughout the project. Again, minimum standards must be met on validity packets to continue to work on the project.

Continued Monitoring of Sites

MI's Client Command Center/Project Command Center software program allowed MI Scoring Directors and Project Management and MSDE to view daily and cumulative reports on score point distribution, agreement rate between Scorers on second readings, and numbers of responses scored. These reports were arranged by item, and information could be accessed for an individual, team, or the entire group for a specific content area.

Each Scoring Director submitted daily progress reports to the MI Project Director. These reports detailed activities during training and scoring, noting any problems or delays encountered. Project Management also communicated with the Site Managers, Project

Monitors, and the Scoring Directors via email, phone, or fax, or by visiting the scoring centers, as needed.

Scanning and Data Reporting

All scanning of score sheets took place at the scoring center where the reading took place. Scores from these scans were transmitted via secure connection to MI's Information Technology Department in Durham, where the information was merged into data files for each form. These data files also included the scores from the machine-scannable selected and gridded response items, which were scored by machine in Durham during the initial processing of student booklets. After these data files were edited for accuracy and completeness, they were securely transmitted to CTB/McGraw Hill and MSDE.

Conclusion

The size, scope, and scale of the May 2001 Administration of the Maryland High School Assessment was a challenge to all of those involved. Multiple content areas, multiple forms, and multiple items per form—scored at multiple locations—add up to a logistically complex project. MI is currently building its capabilities in our IT and processing operations to continue to meet that logistical challenge. We are currently reorganizing our IT Department for greater efficiency and increasing its resources, both in manpower and facilities. MI has been in contact with MSDE and CTB to streamline the security check in process.

Our scoring personnel are also preparing for a project that will increase in complexity as the focus changes from field testing to individual student accountability. Training procedures currently in place will not need to be changed substantially, except that qualifying sets and validity sets will become additional tools for training Scorers and monitoring their performance. To allow for possible dismissals or other attrition related to these new standards, in addition to allowing for 90% more second readings, Team Leader and Scorer staffing of the project will need to increase proportionally.

Measurement Incorporated's Maryland High School Assessment staff has grown in size and experience. It will continue to do so to meet the challenges ahead as the assessment grows and evolves.