Maryland High School Assessment

2002

Scoring Contractor's Report

Measurement Incorporated
423 Morris St
Durham NC 27701

October 2002

Introduction

The 2002 Maryland High School Assessment consisted of two test administrations, the January Field Test and the May Pre-Operational Assessment. Both administrations were composed of multiple forms, each form with multiple constructed and selected response items for each of the five content areas assessed: Algebra/Data Analysis, Biology, English, Geometry, and Government. One form in each content area (two forms in the case of Algebra) were "repeat forms" administered and scored in both January and May.

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, 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 Content Area Scoring Directors for each specific content area. A guide and practice sets were created for each item in the assessment and included annotations linking the rubric to the specifics of the student responses, thus providing the rationale for the scores.

Additionally, qualifying and validity sets were included this year in order to ensure better quality control of the accuracy of scores assigned to constructed response items. Since the High School Assessment now includes reporting of individual scores for each student, quality control of scoring is an even more important issue. One form in each content area was selected by MSDE as the "qualifying form." Immediately after training, Team Leaders and Scorers had to meet a minimum standard of agreement (set by MSDE) when scoring these qualifying sets in order to continue to work on the project. Two validity sets, similar in design to the qualifying sets, were created for all forms. These sets were administered to each Scorer at least once per week of scoring to ensure that Scorers were continuing to assign accurate scores based on Maryland criteria.

Professional Scorers who had been systematically trained by each Content Area Scoring Director using the above-mentioned materials scored each test. 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, approximately 45,000 answer books were hand-scored in January, and data was collected from nearly 100,000 individual score sheets. Over 265,000 answer books were handscored in May, and data was collected from over one million individual score sheets.

2002 also marked the first year that all student responses were required to receive double readings. In the past, 10% of all responses received a second reading for the purposes of quality control. For the January administration, Algebra/Data Analysis, Biology, English, and Government constructed items

received 100% double readings, while Geometry was designated to continue with 10% second readings. For the May assessment, Geometry also received 100% double readings (with the exception of field test items repeated from the January form, which again received 10% double readings for statistical purposes).

Additionally, the MI Content Area Scoring Directors performed third readings in the rare cases that the first and second reading scores were not at least numerically adjacent in agreement (e.g., a score of "one" and a score of "three"). Third reading scores assigned by the MI Content Area Scoring Directors overrode the scores of the first and second scorers. First and second reading scores that agreed perfectly or that were adjacent were both reported in the raw score data sent to MSDE and the Development Contractor.

Third readings also were given to responses that received a score of "zero" in conjunction with any other score. As with non-adjacent scores, the third reading score overrode those of the first and second scorers. This procedure was included to ensure that all responses receiving a score of "zero" were "completely incorrect or irrelevant" as the rubrics require for that score. This same process applied to responses that received a condition code and any other score.

With scores being reported at the student level, the importance of assigning the correct scores to student responses was paramount. Double readings with resolution helped to ensure the accuracy of the raw holistic scores assigned to the constructed item responses by MI scoring personnel.

As an adjunct to the statistical evaluation of items, Scorers used an item evaluation sheet daily to record trends and idiosyncrasies observed when scoring student responses. Each Team Leader reviewed and summarized each team's comments, adding his or her own as well. Each Content Area Scoring Director discussed each item with the Scorers, read the evaluation sheets and Team Leader summaries, added his/her own observations, and compiled an anecdotal report on scoring for each item. MI Project Management reviewed this item evaluation information and supplied it to MSDE after the scoring of each test administration.

In every aspect of the 2002 HSA scoring conducted by MI, there was a collaborative effort with key staff in the MSDE. The MSDE Director of Scoring and the MSDE Content Specialists were consulted for all decisions, creating the scoring process that Maryland directed and that MI executed.

Staffing

Scoring Project Management

The function of MI Scoring Project Management is to coordinate and execute all handscoring and related activities for the project. The MI Project Director and the Assistant Project Director work closely with MSDE content and scoring

personnel, act as liaisons between MSDE and the MI Content Area Scoring Directors, and, through MSDE, coordinate activities with other contractors. The Project Director and Assistant Project Director oversaw all MI Content Area Scoring Directors, Team Leaders, Clerical Aides, and Data Processing staff. Scoring Project Management also was responsible for overseeing day-to-day management at all scoring facilities where HSA scoring took place and for the development of all scoring guides and other training materials, as well as all the materials used to maintain quality control in training and scoring. Scoring Project Management was also responsible for the training of MI Content Area Scoring Directors.

Additionally, our on-site Project Monitors continued to be a valuable and important part of MDHSA scoring project. Project Monitors oversee and administer all scoring projects assigned to their scoring site and communicate daily with both MI Content Area Scoring Directors and the Project Director and Assistant Project Director.

Content Area Content Area Content Area Scoring Directors

MI staff for the HSA expanded to meet the need for a growing program. This year, MI bolstered its Content Area Scoring Director staff for the Maryland High School Assessment. Since each content area is scored by at least two different groups of Scorers (e.g. BCR and ECR), it was necessary to add additional Content Area Scoring Directors to the project in order to accomplish training and scoring with the efficiency and alacrity required.

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

MI Content Area Scoring Directors were diligent in adherence to HSA scoring standards and ensured that Team Leaders and Scorers assigned scores to student responses based on these scoring standards. While they competently addressed scoring issues unique to their content areas, they also recognized issues for which precedent has not been established. They presented these issues to MI Project Management, who conferred with MSDE scoring and content specialists for guidance and resolution.

Team Leaders

In selecting HSA Team Leaders, MI's management staff and the Content Area 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 the HSA or similar projects, as well as HSA 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 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 12 Scorers, each Content Area Scoring Director had a floating Team Leader. This person directly assisted the Content Area Scoring Director in maintaining paper flow and supervising Team Leaders, and helped 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. As well as employing many of our experienced Scorers for this project, we also recruited new ones.

Each scoring site recruited new Scorers for this growing project. 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.

Site Managers

- MI used multiple scoring sites in order to accomplish the large task of scoring the High School Assessment. Each MI 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

supervised on-site warehouse and clerical personnel involved in the scoring project.

Rangefinding

Overview of the Process

The rangefinding process is the first and most important step in the handscoring of constructed test items. Constructed responses are an integral part of Maryland's High School Assessment Program, providing student-produced evidence of application and reasoning as valued in Maryland's educational programs and strategies. Collaboration between the Scoring Contractor, MSDE Testing and Content Specialists, and Maryland educators is the MSDE's cornerstone for the successful scoring of the Maryland HSA program.

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 referred to as rangefinding. This process is pivotal to the success of Maryland's testing program, calling upon the expertise of Maryland educators in concert with the scoring contractor's professional staff. It is the foundation of constructed response scoring.

Preparation for Rangefinding

The same day that each test was administered a sample shipment of each completed test from schools selected by MSDE was express shipped from the school to the Measurement Incorporated Central Office in Durham. This is referred to as an "early delivery sample." The responses in these tests were carefully reviewed by MI specialists, in accordance with the generic rubrics and anchor sets, who selected a variety of responses for Maryland educators to evaluate. In 2002, much of this work was done by the MI Content Area Scoring Directors and Team Leaders at the scoring sites, rather than being centralized in Durham. This process allowed for rapid and accurate selection of responses by personnel with extensive HSA scoring experience. The selected responses were assembled in packets that contained an adequate number of responses to show the full range of the early delivery sample and a variety of student approaches to each test item. These responses were duplicated to provide a copy for each committee member.

The assumption is that the early delivery sample will be representative of the whole assessment. However, whenever a new student approach to a response occurs during the actual scoring, MI always consults the MSDE Director of Scoring and the MSDE Content Specialists for direction. MI is diligent in implementing Maryland decisions 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.

Rangefinding Meetings

Committees composed of educators from Maryland schools and MSDE, along with MI Project Directors, Project Monitors, and Content Area Scoring Directors, met prior to the January and May 2002 scoring of constructed responses to pre-score a sample of 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. The 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 of Maryland educators.

After each committee as a whole was trained and calibrated on previously scored HSA test items (anchor 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, while Biology separated into two sub-groups for BCRs. (This was necessary because of the high number of Constructed Response items in these two content areas—7-9 BCRs per form for Biology, and 8 BCRs and 1 ECR per form for Government.)

2002 marked the first year that qualifying sets and validity sets were included in the training and monitoring process. This meant that a larger number of sample responses had to be scored for each item by the rangefinding committees than in previous years.

January Rangefinding Conference

MSDE Scoring expressed interest in holding some January rangefinding meetings at the MI Scoring Centers involved in MSDE scoring. This plan gave MSDE personnel the opportunity to become familiar with the scoring centers where the HSA is actually scored. It also fulfilled MI's desire to increase the involvement of our satellite centers in the rangefinding process.

Content Area	Dates
Algebra	January 23 - January 26
Biology	January 21 - January 25
English	January 28 - February 1
Geometry	January 29 - January 31
Government	January 24 - January 26; January 28 - January 29

Team Leader and Scorer Training

Preparation of materials

Upon the completion of rangefinding, MI Content Area 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. Additionally, 2002 marked the first year that qualifying sets and validity sets were included in the training and monitoring process.

One guide and two training sets were created for each item. Guides typically consisted of three to four anchor papers per score point. More examples of each score point were included if a corresponding variety of types of responses were found in rangefinding. The number of sample responses for each item varied not only with the complexity of the responses and the extent of the score scale, but also with the variety of student approaches to the item as encountered in rangefinding. Guides included rubrics, annotated anchor papers for each score point, and scoring guidelines for each item. Examples of responses at each scorepoint were included in the scoring guide in scorepoint order with annotations to link the rubric to the specifics of the student response, thus providing the rationale for the score.

.

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.

HSA testing currently consists of multiple forms with unique CR items per content area. To make the training and qualifying process more practical, equitable, and efficient, MI and the MSDE scoring staff worked together to develop a training protocol, first used in January 2002 scoring, to allow for Scorer qualification based on performance on qualifying sets in training. One form for each content area was designated by MSDE scoring as the “qualifying form” for that content area. Qualifying sets consisted of approximately 20 -25 responses each, including all items in the qualifying form item group, BCR or ECR.

After completing training on the guides and training sets for each item, each Scorer then completed at least two qualifying sets and had to achieve a minimum standard of perfect agreement with the true scores (consensus scores assigned to the responses by the rangefinding committee). Additionally for the qualifying form and for each additional form, validity sets were created. These sets, identical to the qualifying form in structure, were given to each Scorer at least once per week in order to ensure that the Scorer was still assigning accurate scores based on Maryland's criteria.

Any changes in training materials that became necessary as the project evolved were completed with approval of the MSDE scoring and content personnel and any such changes were documented. This included decision papers, which were documented with the MSDE decision and date. Copies of each scoring guide and each training and validity set (with answer keys) are provided to MSDE. MI also maintains archived copies of the completed training materials, including annotations.

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

Team Leader Training

After the guide, training, qualifying, and validity papers had been identified, finalized, and approved, Team Leader training began for the first form in each content area. The Content Area Scoring Directors and/or Assistant Content Area Scoring Directors conducted the training of the Team Leaders. Procedures were similar to those for training Scorers (see below) but were slightly more comprehensive, dealing with resolution of discrepant scores, identification of nonscorable responses, unusual prompt treatment (including ESL and dialect), alert situation responses (e.g., child-in-danger), and other duties performed only by Team Leaders. Team Leaders were required to take careful notes on the training papers in preparation for discussion with the Scorers, and the Content Area Scoring Director counseled Team Leaders on training techniques and application of the rubric.

Scorer Training

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, developed the flexibility needed to deal with a variety of responses and retained the consistency needed to score all papers accurately. In addition to the initial scoring training, a significant amount of time was 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.

After Team Leader training and qualifying was completed, the Content Area Scoring Director conducted the training of Scorers. All 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 Scorer was assigned an individual number for easy identification of his or her scoring work throughout the scoring session.

After the contracts and nondisclosure forms were signed and the introductory remarks given, training began. The Content Area Scoring Director presented the constructed-response item and introduced the guide, then discussed, room wide, each score point and example response. This presentation was followed by practice scoring on the training sets. Each Scorer worked individually to assign scores to the responses in these sets.

Team Leaders collected the monitor sheets after the scoring of each training set and recorded results in a customized log which was examined by the Content Area Scoring Director to determine which papers were giving Scorers difficulty. Because it is easy in a large group to overlook a shy Scorer who may be having difficulty, Scorers break into teams to score and discuss the papers in the training sets. This gives Scorers an opportunity to discuss any possible points of confusion or problems in understanding the criteria.

The Content Area Scoring Director also "floated" from team to team, listening to the Team Leaders' explanations and, when necessary, adding additional information. If a particular paper or type of paper seemed to be causing difficulty across teams, the problem was discussed room wide to ensure that everyone heard the same explanation.

Qualifying

Team Leaders and Scorers were required to demonstrate their abilities to score accurately by attaining at least the agreement percentage established by the MSDE before they were allowed to read packets of actual papers. Any Team Leader or Scorer unable to meet the standards set by the MSDE was dismissed. All Team Leaders and Scorers understood this stipulation when they were hired. After reviewing the guide and completing two training sets for each item, each Team Leader and Scorer then completed two qualifying sets, which incorporated items from the cluster of items for that form. In order to continue to work on the project, each Team Leader and Scorer had to achieve a minimum percentage of agreement with the "true scores" assigned by Maryland rangefinders to each response in the qualifying set.

Qualifying scores, set for January and May 2002 scoring, currently are tentative and subject to change as needed as the project evolves:

2002 Minimum Agreement Rates for Qualifying

Content	Agreement
Algebra	80%
Biology	70%
English	70%
Geometry*	80%
Government	70%

*for May assessment only

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 (Government ECR and English ECR had only one item per form), and a separate training session was held for each new set of items to be scored. Each training session for additional forms was conducted in the same manner as the initial Team Leader and Scorer training sessions, except that qualifying sets were not included.

Handscoring

Overview

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 scan-able 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 by team to the Scorers. All of the packets were read twice. (Geometry received only 10% second readings in January and 100% in May, with the exception of January field test items.) These packets contained two score sheets, one for each reading. Also, the second Scorer used a separate score sheet and was unaware of the scores assigned by the first Scorer. Special care was taken to ensure that the packets identified for second reading were distributed equally among the entire pool of Scorers. No second reading packets were distributed to the same team of the Scorer who did the first reading.

As a Scorer completed a packet of papers, he or she placed it back in the envelope and returned the packet, along with the score sheet, to the Team Leader. 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 Content Area Scoring Director for rescoring. 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.

Quality Control of Handscoring

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 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 Content Area Scoring Directors, Team Leaders, and Scorers.

Third, all Content Area Scoring Directors, all Team Leaders, and usually most of the Scorers at MI's current facilities have had previous experience on HSA and/or 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 constantly was monitored during the project:

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

All constructed response items received a second reading. (Geometry items in January and Geometry repeat field test items in May received 10% second readings.) By matching these 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 Content Area 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 entire room. In the case of a two-point disagreement in scores, a third (resolution) reading was done by the Content Area Scoring Director to ensure the accuracy of the score assigned to the response. Third readings also were done for responses that received a score of zero or a condition code assigned with any other score.

MI's Client Command Center/Project Command Center software program allowed MI Content Area Scoring Directors and Project Management and MSDE to view daily and cumulative reports on score point distribution, agreement rates between Scorers, 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.

Three reports are generally used to monitor scoring performance:

- The Inter Rater Reliability Report lists the number of responses scored and the number of those that have been read twice. It indicates how one Scorer's scores compare with the scores from the other Scorer. The result, by percentage, can be Equal (the scores agree), Adjacent High or Low (the scores do not agree, but are adjacent), or High or Low (the scores do not agree and are not adjacent).

- The Inter Rater Split Report is a more detailed version of the Inter Rater Reliability Report used to identify specific scoring trends in individual Scorers. The total number of responses scored and, of those, the number that are second readings are listed. Also, it gives perfect agreement percentages and adjacent agreement percentages and provides the total number of responses "missed" on each side of the scoring line for each score point.

- The Score Distribution Report shows the percentage of responses,

by item, that received a particular score.

Validity Sets

Content Area Scoring Directors selected approximately 50 papers per form per content area that were placed into two unique validity packets of approximately 25 papers each. These were distributed to each team and administered daily on a rotating basis. Each Scorer scored at least one of these packets during each week of scoring. Scorers who were minimally successful in training were the first to be given validity packets. Validity score reports indicated the percentage of papers scored correctly by each Scorer and the number of papers scored too high or too low.

Because the assessment consisted of multiple forms with unique items in each form, the validity scores only indicated how the reader performed as far as the particular items in the particular form being scored. While some scorers did well across forms, others did better with certain forms or items. When a scorer's validity scores were consistently low across multiple forms, that was indicative of a more serious problem in applying the rubric criteria to student responses. MI looks forward to working with MSDE to establish standards for validity scores similar to the current standards for qualifying.

Retraining

Spot-checking, validity scores, and status reports provided project management with continuous feedback not only on individual Scorers but also on room-wide scoring trends. Content Area Area 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 the 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 particular elements of the papers and, therefore, 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 this is rather time-consuming, the "reading partner" strategy generally was 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.

Monitoring

Each Content Area 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 Content Area Scoring Directors via email, phone, or fax, or by visiting the scoring centers, as needed.

Decisions and Alerts

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 Content Area Scoring Directors. After a brief review, project management then forwarded these responses to MSDE scoring and MSDE Content Specialists for scoring decisions. These decisions and the accompanying explanations from MSDE then were given to the Content Area Scoring Directors. In this way, responses with new and unanticipated approaches 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 Leaders or to the Content Area Scoring Directors. At that point, the Content Area 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 Martin Kehe, MSDE, for follow up.

January 2002 HSA forms and constructed item groups

Content Area	Form	BCR Group A	BCR Group B	ECR Group C
ALGEBRA	**S**	**12, 22, 31, 41**		**7, 17, 36, 45**
	T	**16 , 36, 54**		**7, 21,47**
BIOLOGY	**Q**	**8, 13, 17, 23, 29**	**45, 49, 58, 64**	
	R	**10, 21, 43**	**53, 61, 67, 77**	
	S	**7, 14, 21, 26, 31**	**41, 47, 53, 62**	
ENGLISH	**S**	**5, 35, 65**		**47**
	T	**14, 61**		**54**
	U	**6, 14, 49**		**41**
GEOMETRY	**T**	**10, 20, 40**		**5, 15, 31, 44**
	U	**19, 39**		**10, 22, 46**
GOVERNMENT	**Q**	**6, 14, 20, 34**	**46, 52, 58, 71**	**28**
	R	**6, 14, 20, 34**	**46, 56, 65, 76**	**28**
	S	**6, 14, 20, 34**	**46, 56, 65, 76**	**28**

EACH CONTENT AREA ALSO INCLUDED TWO ADDITIONAL MAKE-UP FORMS FOR EACH TEST ADMINISTRATION.

January Administration Scoring

Content Area	Number of Students Scored	Number of Scorers/ Team Leaders	Dates of Activity (from TL training to end of scoring)
Algebra	13,086	BCR group A: 26/4 ECR group C: 21/3	February 4 Through February 27
Biology	21,094	BCR group A: 39/4 BCR group B: 38/4	February 4 Through March 8
English	19,740	BCR group A: 26/5 ECR group C: 16/4	February 7 Through March 6
Geometry	11,000	BCR group A: 10/2 ECR group C: 14/3	February 7 Through February 26
Government	22,828	BCR group A: 31 / 4 BCR group B: 41 / 5 ECR group C: 18 / 3	February 7 Through March 8

For all content areas except Geometry, this administration marked the first time responses received double readings with resolution and the first time that qualifying and validity functions were included. Also, this marked the first time that a separate Team Leader training was held prior to Scorer training.

As noted earlier, zero-one disagreements were flagged programmatically for a third reading. The large number of zero-one resolutions required in Biology and in Government, compared to the other three content areas, may point to a zero-one rubric "gap". That is, there are responses that, while not "completely incorrect or irrelevant", do not display evidence of "some understanding" (Biology) or "minimal understanding" (Government) for a true score of "one".

Adjacent split scores at other rubric levels, score points one and two for example, are valid because there are responses that are truly "line calls" that fall between

the scoring parameters for either score point. Since the zero score point is flatly defined as completely incorrect or irrelevant, a zero-one adjacent score would seem invalid. However, many very minimal responses that include one small piece of correct information do not appear to include enough correct information to display “some” or “minimal” understanding, or to fulfill other rubric requirements of a score point one.

May 2002 HSA forms and constructed item groups

Content Area	Form	BCR Group A	BCR Group B	ECR Group C
ALGEBRA	S*	12, 22, 31, 41		7, 17, 36, 45
	T*	16, 36, 54		7, 21, 47
	U	16, 36, 43, 54		7, 23, 48
	V	6, 17, 30, 41		12, 21, 36, 45
	W	6, 17, 30, 41		12, 21, 36, 45
	X	6, 17, 30, 41		12, 21, 36, 45
BIOLOGY	Q*	8, 13, 17, 23, 29	45, 49, 58, 64	
	T	12, 24, 35, 44	56, 69, 77	
	U	6, 16, 21, 28,	32, 39, 46, 52, 61	
	V	5, 12, 21, 26, 32	39, 50, 56, 64	
	W	5, 11, 17, 24	32, 42, 50, 57, 64	
ENGLISH	S*	5, 35, 65		47
	V	7, 66		54
	W	8, 16, 63		47
	X	20, 37, 61		53
	Y	14, 34, 65		52
GEOMETRY	T*	10, 20, 40		5, 15, 31, 44
	V	7, 36		15, 21, 42
	W	6, 17, 40		12, 21, 31, 44
GOVERNMENT	Q*	6, 14, 20, 34	46, 52, 58, 71	28
	T	6, 14, 20, 34	46, 52, 57, 71	28
	U	6, 14, 20, 34	46, 52, 58, 71	28
	V	6, 14, 20, 34	46, 52, 58, 71	28
	W	6, 14, 20, 34	46, 52, 58, 71	28
	X	6, 14, 20, 34	46, 52, 58, 71	28

*= REPEAT OF JANUARY 2002 FORM

EACH CONTENT AREA ALSO INCLUDED TWO ADDITIONAL MAKE-UP FORMS FOR EACH TEST ADMINISTRATION.

May Administration Scoring

Content Area	Number of Students Scored	Number of Scorers/ Team Leaders	Dates of Activity (from TL training to end of scoring)
Algebra Charlotte	128,936	BCR group A: 56/5 ECR group C: 72/7	June 10 Through July 30
Biology Nashville	101,114	BCR group A: 51/6 BCR group B: 41/7	June 11 Through August 9
English	106,116	BCR group A: 65/9 ECR group C: 39/4	June 19 Through July 24
Geometry	91,746	BCR group A: 30/4 ECR group C: 31/4	June 17 Through August 7
Government	105,264	BCR group A: 64/8 BCR group B: 71/9 ECR group C: 18 / 3	June 10 Through August 2

In all content areas, including Geometry, all responses were double-scored with resolution readings. (Geometry field test items repeated from January received only 10% second readings.)

For Biology, it was necessary to train and qualify an additional group of scorers (a “wave”) after the initial group had begun scoring. The wave was trained using the same guides, training sets, and qualifying sets as were the original group. Assistant Content Area Scoring Directors monitored the original group while the Content Area Scoring Directors trained the wave. After qualifying, the wave then scored Make-up Form 1, which had the same items as Form Q, the qualifying form for Biology. The wave then joined the original group for training on and scoring additional forms.

There continued to be a large number of zero-one resolutions required in Biology and in Government compared to the other three content areas. The large number of students tested in May meant that the number of resolution readings increased

substantially over those in the January assessment. This factor created a large backlog of answer documents that required resolution readings.

Materials arriving late from the schools created delays in reporting scores for all forms in all content areas. Late shipments, some of them quite substantial, arrived from several schools long after the fourth pick-up was made at the schools. These materials were processed and scored as soon as possible upon their arrival. Some forms in some content areas required supplemental score files sent after scores had been reported to the development contractor.

Materials Handling and Data Reporting

Pick-up and Transfer of Test Materials

MI arranged for a Maryland-based courier service to handle the pick-up of the HSA test booklets and established a schedule for the four separate pick-ups that were needed. The first three pick-ups included all of the schools involved in the administration of the test. The fourth and final pick-up included each LEA's central office in addition to all the school pick-ups.

Upon each pick-up, school personnel received from the courier a receipt listing the actual number of boxes. After each pick-up of materials from the schools and the LEA offices, the materials were not transferred between the courier service's trucks until they reached the courier's central facility in Maryland. This procedure contributed to the security of the project because reducing the number of times each school's materials were handled also reduces the possibility of any materials being misplaced.

Upon receipt at the courier's central facility, the materials once again were verified against the database as to quantity, subject, and school. After all of the pick-ups were completed for that day, the materials were packed for shipping directly to our main offices in Durham, North Carolina. MI arranged for a national shipping company to provide this service. In order to maintain tight security of the materials involved, this shipper was required to take the test materials from the the courier's central facility directly to our receiving center in Durham. There were no stops for any other pick-ups or at any warehouses operated by the contracted carrier. As an extra security measure, there were no other materials on the delivery vehicle(s) except for the test materials to be delivered directly to Durham.

We experienced a number of problems this year, especially this summer, with our courier and the pick-up process. Most of these problems seemed to be due to the large size and scale of the task (combined with MSPAP pick-ups). MI is working with the courier and their subcontractors to correct these problems and to prevent a reoccurance.

Serving as a vendor directly for MI and working from our predetermined schedule, the courier service will be required as the four schedule pick-ups are made to give MI immediate feedback when any problems arise. They will contact us regarding delays due to traffic, weather, etc. Our personnel will always have direct telephone access to the courier's supervisors during pick-ups so that any problems can be resolved immediately. Additionally, all couriers will be required to have appropriate identification for themselves and for their vehicles, which will all have enclosed cargo areas.

Despite the problems experienced in working with the courier this past May, we believe that using a Maryland-based courier service offers advantages over a large parcel pick-up services (i.e., UPS, FedEx, or Overnight Express) First of all, revenue is recycled back into the local Maryland economy. Second, using a Maryland courier greatly increases our ability to track all of the materials involved, thus enhancing our ability to maintain the security called for by this project.

Check-in and Processing of Materials

Upon receipt of the test materials in Durham, the first priority was to match the security codes of returned material with the material originally shipped. This was done in two ways. The security bar codes on the test books for all subject areas and the security bar codes for unused mathematics answer documents were scanned by warehouse personnel using a handheld or flatbed scanner. Meanwhile, the used answer documents were at the scanning center, where the used mathematics answer documents (the only content area with secured answer documents) had their security bar codes read by our OpScan scanners. The bar code data from both scanning processes was downloaded into specific programs to be compared with the data from the original shipment.

MI followed a number of quality control, back up, and identification procedures during the security check-in process. Each box opened in our warehouse was assigned and tagged with a unique school and data bar code identifier, so if any questions about certain bar codes arose, the original documents could be located quickly. Also, all scanner files were backed up daily, or more frequently, depending on the volume being processed.

The boxes of materials received at the MI warehouse contained test books and answer documents. The used answer documents were sorted by form and placed in boxes for scanning. The boxes were then taken to our production area for scanning and processing. As the test books were loaded into the scanner, our scanning personnel assigned a unique batch ID number to each box, and a batch number label was attached to each box. This label also identified the content area and the form contained in the box. After scanning, the tests were put back in the labeled boxes and sent on to our packet making staff. Only one form was put into a box, and boxes were always maintained as separate units. The

answer document covers remained in the boxes after the CR item pages have been split apart, arranged in packets, and sent to scoring.

Collection of the bubbled information is controlled by a computer program that tells the scanner where the bar codes, litho codes, and bubbled information areas should be located on each answer sheet. Each of these areas is mapped to a definition that specifies what data is valid for that area of the answer sheet. The information recorded from a single answer book, including student identification information, litho codes, and selected responses and gridded response items, is represented as a line of data, or record, in a text file. Additional information assigned during the scanning process also is present in each record, such as the identity of the scanning program, the batch number, a unique sequence number for each document scanned, and the scanning date. Each line of data in the file represents a different answer book, and each file contains information only from the books of a single batch, which represents a physical box.

Upon completion of handscoring, raw scores assigned to CR items were merged with the data collected from the scanning process using the same computer program that initially generated and assigned packet numbers and packet positions to the student books. Since these numbers do not rely on any scanned data, they are an extremely reliable means to ensure that each handscored data record is correctly matched to its student data file from the scanning of the actual answer documents. Careful attention was given to having 100% complete matching of demographic information with associated SR and CR items.

Final data files were generated from the master database server. These data files were made available to the Test Development Contractor in the desired format. In addition, the final files uploaded to the MSDE server were processed through a quality assurance system developed by our IT personnel. Each column of data was analyzed based on the type of data valid for that column. The validation requirements were derived from the file layout and descriptions provided by the Test Development Contractor when they initially transferred the student data files to our database server.

Any questionable data was verified by examining the original data files and/or the original answer document or score sheet. The quality assurance system is, in actuality, a double check, because the definition information provided already has been applied to each data field by the scanning data validation processes prior to the information being stored in the project's master database.

Problems
A number of difficulties affected the production of the security check-in report.

- Student barcode labels or other types labels covered the security label.
- Secure materials sometimes had no readable security barcode.
- Biology and Make-up materials were not included in the CTB data.
- Large Print Booklets consistently were labeled with product codes and other type barcode labels that were placed over security barcodes printed on the physical documents. None of the various barcodes appeared in the data from the Development Contractor.

Storage of Materials
As the Scoring Contractor, MI will store all test books, used answer documents, and unused math answer documents for the entire contract period. When an entire pallet of storage boxes containing test books was completed, a pallet inventory was produced, detailing the unique bar code numbers of the boxes as well as descriptions of the boxes' contents. This clearly identifies materials for storage, retrieval, and eventual recycling. Answer documents were filed in packet order and labeled before being placed in storage. All materials were stored such that retrieval and shipment to Maryland of any documents requested can be accomplished within a 24-hour time frame. After the contract has been completed, MI will await further directions from MSDE as to the disposition of these materials. If MSDE advises that the materials should be recycled, all test books and all unused mathematics answer books will be recycled in a secure manner. All unused answer books also will be recycled.