Maryland High School Assessment
May 2001 Field Test
Analysis Specifications

Version 1.3

Overview

The purpose of these specifications is to describe the research analysis procedures and deliverables for the May 2001 Maryland High School Assessment (MD HSA).

Background Information

Beginning in 2004, MD HSA is scheduled to become a high-stakes test used as one factor in determining high school graduation. The MD HSA tests are in Algebra, Biology, English I, Geometry, and Government. The tests consist of mixed-item-formats, including selected response (SR), brief constructed response (BCR), extended constructed response (ECR) and, for Algebra and Geometry, student produced response (SPR; also known as grid response items [GR]).

Field tests will be administered in Maryland in January and May from 2000 to 2003. Each form will look like future operational forms in terms of the numbers of items and item types. Beginning in 2001, all eligible students in the state will participate in the field tests. School summary reports will be produced for all students completing the pre-operational form as defined below.

For administrations from 2001 - 2003 there will be 2 types of forms administered to the students. The first type of form is a "pre-operational form" comprised of previously field-tested items and 30 minutes worth of field test items. This form meets the content blueprint and specifications for future operational forms. In addition, block field test forms will be administered that also follow the operational field test blueprints. One block field test form will contain the anchor item set (25-30 items), the other block field test forms will contain an additional 30 minutes of field test items (SRs, CRs, or a combination of both item types). Summary scores will be reported for students based on the pre-operational form.

The pre-operational and block field test forms will be spiraled at the classroom level. This was achieved by stacking the test versions in alternating fashion so that they would be distributed at equal rates within each classroom across the state. For example, six Algebra versions were administered 1-2-3-4-5-6-1-2-3-4-5-6 and so on to individual students within each classroom. Such an approach has two benefits: (1) approximately the same number of students from throughout the state would be administered each version, and (2) randomly equivalent groups of students in terms of socioeconomic background and achievement levels would take each version. This equivalence of samples (Random Groups Design) taking each form is to help ensure that the items will be accurately placed on the score scale.

Anchor item sets were included in one of the block field test forms to help ensure the successful placement of item statistics onto the MD HSA field test scale (based on the May 2000 administration). The anchor item sets were also administered in the 2000 and 2001 forms. Each anchor item will fall in exactly the same location as it appeared in the 2000 administration. Items selected as anchors represented the content of entire test and could be considered a "mini-test". Anchor items are distributed throughout the test.

The current design allows for the scaling of the 2001 field test items using the random group equivalence assumption. Therefore, a single calibration using the test data from randomly equivalent groups of students will be completed. Items will then be placed onto the May 2000 field test scale using the anchor items contained in the block field test book. The following table shows the item type blueprint for the operational forms. Please note blueprints for Algebra and Geometry have changed recently.

	ENGLISH	[illegible]	[illegible]	[illegible]	[illegible]
SR	50	26	26	48	50
BCR	2	3	2	6	7
ECR	1	3	3	1	1
SPR	0	6	6	0	0
TOTALS PER FORM	53	38	37	70	58

2001 Forms

For May, there will be 4-6 forms administered. The table below lists the form identification letters. In all cases the first letter is the pre-operational form, the remaining forms are the block field test forms. The first block field test form contains the anchor set (e.g., for January Algebra, Form L is the pre-operational form, Form M contains the anchor set, and Form N contains all new field test items). The same pre-operational form will be used in January and May.

Form Identification

	English	Biology	Geometry	Govern.	Algebra
Valid January admin. form letters	K, L	J, K	N, P	K, L	L, M, N
Valid May admin. Form letters	K, M, N, P, Q, R	J, L, M, N, P	N,Q, R, S	K, M, N,P	L, M, N, P, Q, R

Note for the May administration Algebra forms M and N will be re-administered due to the low numbers of students completing this form in January. For analyses the data from January and the data from May will need to be combined.

Braille editions of the pre-operational form have been constructed. In some cases (e.g., Geometry and Algebra) items were could not be translated into Braille; therefore, the form may have missing items. This information is required for the computation of student scores as the total possible points for the Braille versions may be different.

The same form (the pre-operational form) is used for Braille in both the January and May administrations.

Braille Forms and items deleted

	English	Biology	Geometry	Government	Algebra
Braille Form letter	K	J	N	K	L
Braille Form Number	10	09	13	10	11
Items deleted?	No	No	30, 37, 54, 55	No	15, 24, 39, 47

Schedule Of Delivery

Estimated timelines for research are noted in Table 3. The dates indicated in the column "Data File available to Research" indicate when the first forms will be available from MI scoring for the calibration.

Table 3. Estimated Timelines for Data Analysis

Content Area	Complete data File available to Research	List of items contributing to scores	DAT files to Development	Board Report Summary to MD HSA
English	August 1	Same as January	Oct 1	Nov 1
Geometry	August 10	Same as January	Oct 1	Nov 1
Biology	August 17	Same as January	Oct 1	Nov 1
Government	August 22	Same as January	Oct 1	Nov 1
Algebra	August 24	September 14*	Oct 1	Nov 1

this date needs to be confirmed with Technology

Data Processing and Quality Assurance

This section specifies network directories, file naming conventions and locations, valid case criteria, data quality control procedures, and rwo, ctl, and anchor file creation.

Directories on Network

The following directory should be used for the MD HSA analyses:

M:\Projects\ mryhsch\????\xxx\TT\CC

where,

mryhsch = Maryland High School Assessment

???? = year
xxx = administration month (Jan=January; May = May)
TT = type of item (FT = field test; OP=operational)
CC = content area (Al=algebra; BG=biology; EN=english; GE=geometry; GV=government)

That is, all analyses for the 2001 MD HSA Algebra May Field Test will be found in the directory:

M:\PROJECTS\mryhsch\2001\May\FT\AL

Naming Convention

The structure of the MSDE naming convention that should be used for the May Field Test analyses will be: CCVMY.*

Where,

CC=the content area (AL=Algebra; BG=Biology; EN=English; GE=Geometry; GV=Government)
V=version of the field test (E=Form E; F=Form F; G=Form G; H=Form H; Z= all forms combined)
M=month of administration (J= January; M=May)
Y=year of adminstration (1=2001; 2=2002)
* = the file type (xls; CTL, RWO, PAR, etc)

For example, the Biology Form P par file from the May 2001 administration would require the following filename: BGPM1.par.

PEID IDT

The IDTs for each content area will be "locked" down and released prior to the receipt of field test data. These files contain specific item information including, book order number, answer keys, item type (SR, SPR, BCR, ECR), item designation (anchor, field test, etc.). These files will be required for scoring student responses, item analyses, control files, RWO-Item Maps, construction of scoring tables, and DAT files. The PEID files released from Paul Calcaterra should be considered the "final" copy. If updates are received, previous versions are to be replaced as soon as the updated version is released.

Copies of the IDTs for each content area will be placed in the PEID folder by Stat Analysis for each administration:

M:\PROJECTS\mryhsch\2001\May\FT\CC\PEID.

Item Summary Information

From the PEID IDT files, Statistical Analysis will produce the following tables (aka Table 2) for each form:

- A summary of the numbers of items by type and designation.
- A listing of each item by designation and the associated standard/benchmark should also be produced.

Please place this information in the following location:

M:\PROJECTS\mryhsch\2001\May\FT\Classical\CC\Table 2.

These tables will be used by research to verify the items contained in the printed test forms.

Valid Case Criteria

After the field test, Measurement Incorporated (MI) will scan the answer documents, score the constructed-response items, and send the complete data files to CTB. Technology will receive the electronic files from MI, ensure that only valid cases are included in the file, and score the data. Each record that arrives from MI will contain the complete information for a single exam, including: all student bubbled information (biographical information, SR raw responses, SPR items raw responses,), all information from the barcode (LEA number and school number) and CR score information (all reads). All forms of all content areas will arrive in the same layout.

An edit program to identify data inconsistencies and incomplete records (i.e., invalid cases) will be completed in Technology. This edit program will produce a report that will be sent to MI for data clean-up purposes. For information, complete descriptions of the project specifications for Technology (Maryland High School 2001 Project Plan and Requirements Specifications for Technology).

Cases will be considered valid and included in the data file based on the criteria listed below. Invalid cases for the criteria denoted with an asterisk (*) should be identified by Technology and not be included in the file that is forwarded to Statistical Analysis. Multiple marks symbol will be the standard CTB multiple mark symbol (a dash “-“). Please verify that all cases received by Statistical Analysis are valid.

Valid Cases
To be valid, cases should have:

- LEA number, or blank
    - School number, blank, or non-participating school
    - An identifiable content area*
    - A valid version (form) and answer sheet*
    - Grade between 7-12, double mark, or blank*
    - Gender M, F, double mark, or blank*
    - Ethnic 1-5, double mark, or blank*
    - Spec Ed Y or N, double mark, or blank*
    - 504 Y or N, double mark, or blank
    - ESL is double mark or blank
    - No out-of-range item response*
        - CR: A – D, or 0 – maximum
        - MC: 0-4, blank, or double mark
        - GR: 0-9, blank, “/”, or double mark
    - Appropriate number of reads for CR items*
- All CR items will receive a single read with a 10% second read. There is to be no resolution of scores. That is, the first read would be the score assigned to the student.

Students with Accommodations

Beginning in 2001, all students completing course work in one of the 5 content areas will be expected to complete the associated examination. There are approximately 37 different accommodations that can be coded for these examinations. It is important to include as many students as possible in the data analysis files, however, in some situations, accommodated students will be excluded. These situations vary by content area. Students with the following accommodations should be excluded from all analyses, including the Red Team Review.

Content Area	Accomm #	Accomm. Name
English	III D	Use of electronic devices (e.g., mechanical speller, computer, augmented communication device, etc.).
English	IV F	Verbatim audiotape of entire test.
English	IV G	Verbatim reading of selected sections of test or vocabulary.
English	IV H	Verbatim reading of entire test.

Data Quality Control Procedures

Scoring Data

The data file will be initially scored by Technology and forwarded to Statistical Analysis. Statistical Analysis should independently score the received data and compare the results for quality assurance (QA) purposes. Frequency distributions for both scores will be produced. Any disagreement between the Technology and Statistical Analysis scores will be investigated. Statistical Analysis will notify the research monitor immediately if data problems are encountered.

SR and SPR items should be scored using the PEID answer key information. Only the Algebra and Geometry versions contain SPR items. The answer grid for all SPRs consists of 5 columns, each containing bubbles for 0-9 and a decimal point. In addition, the three center columns contain bubbles for a division symbol (slash "/"). The parsing rules implemented by, and listed in the Technology Specifications, will be independently verified by Statistical Analysis. For this and future administrations, the parsing rules, the raw SPR responses produced on the GRT, and the scores associated with each SPR will be verified using an independent parsing and scoring program created by Statistical Analysis.

For BCR and ECR items there will be a single read with a 10% second read; however there will be ***no resolution of scores. The first read would be the score assigned.***

Item Analysis and Establishing the Accuracy of Received Data

These procedures follow the standard procedures documented in the "Methods of Establishing the Accuracy of RWO Data" (Yen, Dec 13, 1999).

For each field test version, Statistical Analysis will provide the item analysis output to the Research Monitors in both "hard copy" format and an .xpt file (SAS Viewer 7.0) that is suitable for transfer to an EXCEL spreadsheet. Note that because scores do not count, in some cases student motivation to complete all items may be low; therefore, all item-level analyses (IA) should be completed excluding missing or omitted item responses. For example, p-values and point-biserials will only be calculated based on those students that answered the item. For CR items please note that all students receiving a condition code of ***"A"*** or *"B"* will be considered to have omitted the item. In 2004 it will need to be determined how omitted responses for the item calibrations will be treated. Typically such responses are treated as incorrect or assigned the lowest score in the case of BCR/ER, though speededness issues may require use of trailing omits.

Please obtain for each item:

1. number of students administered the item
2. valid case counts for each item (i.e., all students that answered the item)

3. verify that the logic to include/exclude valid/invalid cases described above has been incorporated
4. a separate frequency distribution of unscored raw data and scored item responses for each version administered to check for out-of-range and/or unexpected item scores as well as determine missing data procedures.
5. standard item analysis that includes:
   - book order number
   - associated goal/standard
   - item designation (anchor, field test, operational)
   - for MC items
     - p-values, standard deviations
     - point biserials for each distractor as well as for correct answer. Note: please exclude the studied item from the total score when calculating the point biserials.
   - for CR items
     - item means, standard deviations
     - frequency distributions and item-test correlations. Note: please exclude the studied item from the total score when calculating the item – total correlations.
   - number and percent of students that did not answer the item (those students that did not reach or omitted the item). For SR and SPR items this will be identified by a blank response. Please note that because we are not independently scoring the SPR items for this administration, the number of students that did not answer SPR items will need to be calculated from the parsed data. For BCR and ECR items these will be all students that have a condition code of A.

As part of the item analyses, please flag items with poor statistics. In particular, flag items with point biserials below 0.15 on the right answer, positive point biserials on one or more distractors, p-values less than .30, or omit/not reached rates greater than 5%.

6. For SPR items please provide an FD of the final edit responses (completed by technology). This information will need to be forwarded to the Math content editors at CTB and MSDE.
7. Separate IA's for anchor item sets for each version of the field test forms (as described in above)

Please include a title/label of the item analysis output that notes the field test form and version (e.g., MD HSA January Field Test Biology).

In addition to the analyses specified above, for each form administered in a content area, Statistical Services will also provide an .xpt (SAS Viewer 7.0) and/or an ASCII .dat file of the following variables in the folder labeled "Raw Data" within each content folder in the directory:

- Form
- Subject ID
- Ethnicity
- Gender
- Grade

- Accommodations Code
- Special Education Codes
- ESL Codes
- LEA
- School
- LEA School combined
- Unscored raw data
- Scored (RWO) data (in the case of CR items this is the first score)
- Condition Codes for each CR item

Please place this file in the appropriate directory for each content area in a subfolder labeled "Raw Data". This data file will be used for future analyses that may be required specific to an individual form and/or for quality assurance checks of data results.

RWO File Specifications

The .RWO files should contain all of the responses for all valid cases across for all forms within the content area. One RWO file will be created for each content area (Algebra, Biology, English Geometry, Government) containing all of the valid response vectors available for all versions of the field test form and a DIF identification code (gender and ethnicity). Please exclude the following cases from the RWO files:

- less than 5 valid responses on the record
- students with the following accommodations:

Content Area	Accomm #	Accomm. Name
English	III D	Use of electronic devices (e.g., mechanical speller, computer, augmented communication device, etc.).
All	IV C	Accessibility to close-caption or video materials.
English	IV F	Verbatim audiotape of entire test.
English	IV G	Verbatim reading of selected sections of test or vocabulary.
English	IV H	Verbatim reading of entire test.

Research will determine and provide a list of items that are to be considered "not reached"via. a trailing omits procedure, if required. These items should be coded as F's. Also, to make it easier to detect any data problems, Statistical Analysis should, wherever possible, retain the district and school building code (these codes are combined to form unique identifiers and are referred to as "LEA Code") for each student in an RWO.

The header for the RWO file should say something like " MD HSA January Field Test Calibration; content area, and date". The data will need to be arranged so that the anchor items from each form are placed at the beginning of the calibration RWO files. Finally, we would like the RWO file sorted by district and school.

Sample RWO file

Anchors	Form L	Form M	Form N	Form P
	XXXXX...XX	FFFF....FFFF	FFFF....FFFFF	FFFF....FFFFF
XXXXX...XXX	FFFF....FFFF	XXX...XXX	FFFF....FFFFF	FFFF....FFFFF
	FFFF....FFFF	FFFF....FFFF	XXXXX...XXX	FFFF....FFFFF
	FFFF....FFFF	FFFF....FFFF	FFFF....FFFFF	XXXXX...XXX

In addition, Statistical Analysis will provide an RWO-Item Map that identifies the RWO position for each item by form, book-order item number, and associated goal/standard. These maps should be labeled following the standard prefix followed by '_RWO_Map' e.g. CCVMY_RWOMap.*
and placed in the following location in the directory:

M:\PROJECTS\mryhsch\2001\May\FT\RWO

CTL Files

Statistical Services will create all of the .RWO and .CTL files and place them in the respective .RWO and .CTL folders in the appropriate network subdirectories prior to the calibrations:

M:\PROJECTS\mryhsch\2001\May\FT\Cal\CC.

One CTL file will be needed for each content area for the calibration. These files will be placed in the folder labeled "CTL" in the directory:

M:\PROJECTS\mryhsch\2001\May\FT\CTL\CC

Items that will be excluded from calibration will be denoted with a "0" in Line 2 of the .CTL file. A SAS program will be used to modify the original CTL files. The modified CTL file will appear in the same folder used in the calibration step. For example if an item is turned off after the first calibration run, the new CTL file will appear in the Version B folder associated with the content area in the cal folder. The SAS program can be located in

M:\PROJECTS\mryhsch\2001\Jan\FT\PEID\Ctl SAS\Invoke ctl.sas

The CTL file for the Windows version of PARDUX requires 2 lines:

Line 1: Sets a variety of parameters for PARDUXMX estimation

- The total number of items to be calibrated
- Maximum number of estimation cycles = 50

- Convergence criterion = 0.001 **(This can be changed to up to .005 if data problems are encountered.)**
- Maximum value of the discrimination parameter (f = 3.4)
- Maximum value of the gamma parameter (7.5)
- Minimum value of the gamma parameter (-7.5)
- Maximum value of theta (θ = 4.10)
- Minimum value of theta (θ = -4.0)
- Number of answer options for the multiple-choice questions (mc item choices = 4)
- Maximum c-parameter (c = .50)
- Estimate 'f' parameter separately for CR item = U
- Bayesian adjustment for c estimates = Y

Line 2: Specifies the number of score levels for each item and is dependent on the item layout for each form calibrated.

- SR items are considered to have one level so "1" is specified .
- SPR items are considered to have two levels (0 or 1) so "2" is specified.
- Constructed response items, for the January sample, have a 4 or 6 point scale. Students receiving a condition code of "A" will be considered to have omitted the item.
- For items that are not to be calibrated a "0" is specified in Line 2. Specific items will be forwarded from Research indicating which items to "turn off" in the CTL file.

Example:

Line 1: 40 50 0.001 3.40 7.50 -7.50 4.10 -4.00 0 0.50 U Y
Line 2: 1114112111111141111111111111111110111111511

The .CTL file for PARDUXMX requires two additional lines:

Line 3: Specifies the number of answer choices for SR items.

- For the MD HSA analyses this will be a 4.
- All SPR and CR items receive a "1" in this line.

Preparation of ANC files

ANC files from the May, 2000 calibrated data will need to be created to link other field-test forms onto the field test scale using the Stocking and Lord procedure. One ANC file is needed for each content area. To create the ANC files, the final parameters for the anchor items from the May, 2000 administration should be obtained and placed into the folder "Anchor files" located at:

M:\PROJECTS\mryhsch\2001\May\FT\Anchor files

Items must be written exactly the same and in the exact order as they appeared in May 2000. Therefore, it is important to compare the anchor items across forms by comparing the actual items in the May 2000 form to the May 2001 form. Any changes in the item (form, content, wording) or location should be noted. In evaluating the equating, any anchor item that was modified or moved will be examined. If the modification or move has produced a significant change in the item statistics, then the item may be deleted from the anchor set.

Note. There is a SAS macro that can be used to obtain these items, however the anchor file was constructed for the January administration and the same file should be used for the May administration. Future administrations can use the same file so long as the anchor items do not change.

The SAS Macro can be located in:

M:\PROJECTS\mryhsch\SAS Macro\AncFiles

Test Analyses

This section describes classical item analyses, IRT calibration and equating, evaluation of results, and DAT file specifications

Classical Item Analysis

This section describes classical item analyses that will be used to conduct the item analysis, review data for reasonableness, and produce the tables that appear in the reports to the Board of Education. To complete these tables the calibration RWO should be used. The summary tables report should be similar to the report produced for the January analyses. Please refer to this as a guide in producing the summary tables for the May administration. The following tables are required for the field test.

Please refer to the following document if you have any questions regarding the layout of these tables.

M:\PROJECTS\mryhsch\2001\May\FT\Specs\Summary Tables for Field Test.doc

Due to the high omit rates associated with several content areas in previous administrations, additional tables need to be provided. Please follow the identical format as the sample tables available at:

M: \PROJECTS\mryhsch\2001\May\FT\Specs\omit rates (aka Speededness)
M: \PROJECTS\mryhsch\2001\May\FT\Specs\Sample CR IA (aka Table 14)

In addition to the Tables included in the Board Report, there are two appendices. The first appendix lists the participant schools and the number of students administered each content area. Stat Analysis will produce this summary in an .xpt file that can be imported into Excel.

The second appendix includes a summary of field test data for each content area. Stat Analysis will provide the following summary data:

- Sample sizes for each version of the form;
- Background characteristics for each form (includes gender, ethnicity, grade, and proportion of sample from Baltimore city);

DIF Analyses

The following procedures will be used to flag items exhibiting DIF:

1. Mantel Haenzel for SR items
2. Mantel for CR items

The flagging criteria and minimal sample sizes for each procedure are listed below:

	MH	Mantel
	SR	CR
Min N per focal group	200	500
Flagging criteria	χ^2_{MH} >0 (p < .05) & $\| \Delta_{MH} \|$ > 1.5	χ^2_M >0 (p < .05) & | ES |>. 25*

Specific information for the calculation of Mantel and ES is available at:

"M:\PROJECTS\mryhsch\2001\May\FT\Specifications\MANTEL.doc".

These will be the criteria that we will use to flag items for review by MSDE. Note, for the Mantel Haenzel and Mantel procedures negative values for the effect sizes favor the reference group, whereas positive values favor the focal group.

The Mantel and Mantel-Haenszel DIF procedures should be completed using the calibration RWO files.

Information submitted to MSDE in the summary tables should include only the field test items. DIF output for anchor and pre-operational items will be provided to MSDE.

IRT Calibration

Overview

These sections describe the analyses involving IRT calibration. If student response is adequate (approximately 1500 responses) the CR items will be included in the calibration for all content areas. Decisions regarding the inclusion of CR items in the calibration will be made following the evaluation of the classical item analyses. The .CTL files will identify which items will be calibrated. Research forward list of items to be excluded from calibration to Statistical Services who will modify the .CTL files accordingly.

Because this is a Random Groups design, random equivalence of the sample is assumed. Therefore, all items will be calibrated together within each content area. Random equivalence will be assumed if:

- The sample sizes for each version are similar;
- Background characteristics for students taking each form are similar (includes gender, ethnicity, grade, and proportion of sample from Baltimore city);

For all calibrations the Windows version of PARDUX, which uses the three parameter logistic (3PL) model for multiple-choice items and the two parameter partial credit (2PPC) model for CR items, will be used. Research will run PARDUX to calibrate the items, review the results to ensure that the correct numbers of items were included in the analyses, the reasonableness of initial estimates, as well as, the Q1 and Q3 statistics for the calibrated items. Items that fail to converge will be identified and hand-estimation will be completed if necessary. After the completion of the hand-estimation, final .PAR files will be created and placed into a **\Final** folder under each content area subdirectory.

Calibration Steps using PARDUX for Windows

Suggested steps to complete the calibration using PARDUX for Windows are outlined below. Save the item summary information for the purposes of checking and evaluating the calibration. The following information should be clearly labeled and saved in the **\Cal** folder for each content area at:

M:\PROJECTS\mryhsch\2001\May\FT\Cal\CC\Version A\

Where: cc = Content Area

a. Estimation Detail/Summary – Pardux Main window text (*.prn)
b. Item Status (*.sta)
c. Distribution – p-values, sample size and level n size. (precalibration *.p and post calibration *.p)
d. Parameters (*.par)
e. SE's (*.abc)
f. Fit Q1 (*.q1)
g. Fit Q3; (*.q3)
h. Estimation Summary (*.sum)
i. Estimation Detail

If more than one version is required, a new folder should be created and labeled appropriately and the above information saved.

1. Start the **PARDUX** Program.
   - When the **File Options** window appears, uncheck the two boxes under **Automatic Read.**
   - Also under **Thetas in input file,** click the **No, any extension** button.
   - Click on OK.
2. Read the control file.
   - In the **PARDUX** window, click on the **File** menu.
   - Click on **Read.**
   - Make sure that the **Files of type** box specifies **Control files (*.CTL).**
   - In the **Open** window, find the appropriate CTL file then click on **Open.**

- Examine the new line appearing in the **PARDUX** window to ensure that the correct control file was read!

3. Read the Response Array File.
   - Repeat step 2 with **Files of type** specified as **Response array files (*.RWO).**
   - Check that the numbers of items and total number of cases are both correct.
   - Note that "Number of items" is the total number of items in the file including excluded items.
4. Check summary statistics.
   a. On the **View** menu click on **reliabilities** in order to check the number of items to be calibrated.
   b. Check that the number of items specified for Alpha and Stratified alpha is the number of items that should be included in the calibration.
   c. Click on **Item Summaries** on the **View** menu and examine item statistics.
      - Click on **Status** and scroll through the items. Items to be calibrated should be listed as **Not estimated** and the correct number of levels should be displayed.
      - Click on **Distribution** and scroll through the items. Item difficulties should be checked for extreme values and OR items should not have zero or near zero frequencies for any level. Save this file to (*.p). This is a good cross check to uncover collapsed levels. Pardux will not estimate levels with fewer than 3 responses.
5. Estimate the item parameters.
   a. On the **Estimation** menu, click on **Parameters**; a **Parameter Estimation** window opens showing the **Estimation Summary** and **Estimation Detail** as the parameters are estimated. Save the estimation summary to *.sum. If C parameters are fixed, this option should be checked in the **Estimation Options** menu.
   b. At the main Pardux Window, save the item parameter estimates to a file
      - On the File menu in the PARDUX window, click on **Save Estimates**.
      - In the **Save As** window for Save as type, specify Parameter files (*.PAR).
      - In the **Save As** window, select the directory in which the file is to be saved.
      - In the **File name** box, type the appropriate file name (e.g., **BGJ01.PAR** for January, 2000 Biology).
   c. Examine the **PARDUX** window (you may need to move the **Parameter Estimation** window to see the **PARDUX** window) to ensure that the estimation procedure converged. A statement like "**Converged at stage 12**" should appear.
   d. If convergence is reached, go to step 10.
   e. If convergence did not occur, there a statement in red font will appear in the main pardux menu e.g. "**Item(s) 120 could not be estimated**".
6. If convergence did not occur, change the convergence criterion up to .005 to see if convergence can occur with a new criterion.
   a. Click on **Estimation** and **Options** increase the convergence criterion to a value up to .005. and click on **OK**. You can determine whether this will help by reviewing the estimation summary to find the lowest g-diff. The **Convergence Criterion can never exceed .005**. Please note on the calibration form changes to the criterion. Maximum estimation cycles should always remain at 50.
   b. On the **Estimation** menu, click on **Parameters** and the parameters will be estimated.

c. If convergence is reached, save the parameter estimates to a file as in 5b above and the new estimation summary as in 5a above.
d. Go to Step 10.

7. If convergence does not occur **(Note: Field Test and Anchor items will not be handfit):**
   a. View the item summary and examine the item statistics
      - In the **PARDUX** window, open the **View** menu and click on **Item Summaries.**
      - Click on **Distribution** and examine the difficulty, total N, and frequencies in each score category (open response items) for the item.
      - Note whether anything appears unusual (extremely easy or difficult item, zero or near zero category frequency).
   b. View the plot of the SR item characteristic curve or the highest category response curve of the CR item.
      - Click on **Thetas** on the **Estimation** menu of the **PARDUX** window.
      - On the **View** menu of the **PARDUX** window, click on **Item Details;** the (full screen) **Item Detail** window appears.
      - On the **Item Detail** menu, click on **Set Active Item.**
      - Enter the number of the item to be examined; a plot of the item characteristic curve (MC item) or the response curve for the highest category (CR item) is displayed.
      - For a CR item, click on the **C/Gamma Level** button; then click on the up and down arrows in the **Active Item** window to examine the different response curves. ***Do not click on the C/Gamma Level button for MC items because clicking on the up and down arrows with that button on will change the c parameter!***
      - Note whether anything appears unusual.
8. Run an M-step on the individual item to see if it can converge without hand fitting.
   a. Click on **Item Detail.**
   b. Click on **M-step** and the results of the M-step will appear in the upper left corner of the window.
   c. Check for convergence; if convergence occurs, save the parameters to a file as in Step 5b above but include a "B" in the prefix to indicate a second version of the parameter file (e.g. BGJ01B.PAR).
   d. Go to step 10.
9. Hand fit an item that did not converge in the M-step.
   a. Click the **C/Gamma Level** button, then click on the up and down arrows to examine the different response curves for a CR.
   b. Save the plots for future reference. For each plot:
      - With the plot on the screen press the **Alt** key and the **Print Screen** key simultaneously. This saves a copy in the windows clipboard.
      - Open a word processor and click on **Paste** to paste the contents of the clipboard into the word processing document (the plot will look best if you format the page in landscape mode). Please save the plots for all hand fitted items in one word document by content area. Identify these by form and item number in the File and create a directory within the content area calibration folder called \Handfit\ Handfitted items Content area.doc. (You should later add the handfitted OPM/EPM plot

c. Examination of the **EPM/OPM plot** for any constructed response item that does not converge, will help in guiding changes, but changes need to be made at the level of the cr item. Click on **Statistics** listed on the **Item Detail** menu. This will display p-values and the current item parameters as shown below.

```
7172 cases:  27796 item total...p=3.8756
à = 1.00000 g[1] =-1.14692 g[2] =-3.46124 g[3] =-5.32234 g[4] =-2.61239
Est m = 0.9693 Actual m = 0.9689
Level      1       2       3       4       5
Obsrv    0.00    0.00    0.01    0.09    0.89
Expct    0.01    0.00    0.00    0.09    0.90
Ob-Ex   -0.00    0.00    0.01   -0.00   -0.00
```

d. The object is to get the Ob-Ex (Observed-Estimated) as close to zero (0) as possible. Keep in mind that as changes are made to the **A/Alpha** and/or **B/Gamma** for one level, statistics for the other levels will change also. A frequency distribution of students' scores (if thetas have not yet been estimated these will be raw-score based) also appears in the plot and is very helpful in making changes to the item parameters. **Remember that every change that is made to one level will affect another. It is more important that the curves fit the data where the frequencies are large than where they are very small. Also, do not make dramatic changes to very low frequency levels. Only make changes that are required to meet the criteria specified above and that look reasonable. Also, adjust levels in decending order by frequency. After changes are made to high frequency levels, adjust the low frequency levels just enough to fit EPM to OPM and appear reasonable in the item detail window.**

e. Click the **A/Alpha** and/or **B/Gamma** button and use the up and down arrows to change the parameters from the default, noting the effect on the fit of the item parameters to the data.

f. When a satisfactory fit is achieved, using the criteria specified above, a set of starting values for item parameters has been specified and the M-step should now be carried out.
   - On the **Item Detail** menu, click on **M-step.**
   - Note whether convergence occurred.
   - If there is convergence, the item will now be considered calibrated. Save the parameters to a file as in Step 5b above but include a "H" in the prefix to indicate that this file includes handfit parameters (e.g. BGJ01H.PAR). Go to Step 10.
   - Save the new plots, which should include level and epm/opm plots for future reference using the steps in 9b, save the parameter file as described in 9e and proceed to Step 10.
   -

10. Examine the Estimation Summary to see if any items were estimated with the maximum *a* parameter. Near the end of the output there will be a line similar to one of the following:

```
0 items with max a; largest a = 2.966
2 items with max a; penultimate a = 2.421
```

If there are any items having the maximum *a* parameter these items should be viewed in detail.

a. Determine the item number:
    - In the **PARDUX** window, open the **View** menu and click on **Item Summaries.**
    - Make sure that the box in the upper right, labeled "Traditional 3PL Metric" is unchecked. This results in the parameterization used by PARDUX in estimation and allows for direct comparison of the *a* parameters with the value in the fourth position in the first line of the control (.CTL) file, usually 3.00.
    - Click on **Parameters** and scroll down until the item(s) having maximum *a* are found.
    - Record the item number(s) so item detail can be examined for this (these) item(s) in later stages.

b. Examine the plots of any item(s) estimated with the maximum *a* parameter.
    - In the **PARDUX** window, open the **View** menu and click on **Item Details**; the (full screen) **Item Detail** window appears.
    - On the **Item Detail** menu, click on **Set Active Item**.
    - Enter the number of the item to be examined; a plot of the item characteristic curve (MC item) or the response curve for the highest category (CR item) is displayed.
    - For a CR item, click on the **C/Gamma Level** button; then click on the up and down arrows in the **Active Item** window to examine the different response curves.
    - ***Do not click on the C/Gamma Level button for MC items because clicking on the up and down arrows with that button on will change the c parameter!***
    - Carefully examine the plots of items estimated with the maximum *a* parameter.
    - If convergence occurred and the ICC plot is not radically different from the data plot do not re-estimating with a larger maximum *a* parameter.
    - If the steepest part of the ICC plot is less steep than the data plot, ***and*** if the overall estimation of parameters did not converge, raising the maximum *a* parameter and re-estimating the parameters ***may*** result in convergence, or at least a solution that better fits the data. Do not use a value greater than 4.00. If it is deemed appropriate to try a larger maximum *a* parameter, modify the control file and begin again at Step 1.

11. Save item summary information for purposes of future checking. **Please not that these naming conventions have changed from that of prior analyses.**
    a. On the **View** menu of the **PARDUX** window, click on **Item summaries.**
    b. Click on **Status**; when the status of each item is displayed, click on **Save** on the **File** menu, choose the appropriate directory and enter a file name (e.g., **BGJ01.sta**).
    c. Click on **Distribution**; when the frequencies are displayed, click on **Save** on the **File** menu, choose the appropriate directory and enter a file name (e.g., **BGJ01pre.p {pre calibration p-values} BGJ01pst.p {post calibration p values}**).
    d. Click on **Parameters**; when the estimates of the item parameters are displayed, click on **Save** on the **File** menu, choose the appropriate directory and enter a file name (e.g., **BGJ01par.txt**).
    e. Click on **SE's**; when the estimated standard errors of the item parameter estimates are displayed, click on **Save** on the **File** menu, choose the appropriate directory and enter a file name (e.g., **BGJ01.abc**).
    f. Click on **Fit Q1**; when the fit statistics are displayed, click on **Save** on the **File** menu, choose the appropriate directory, and enter a file name (e.g., **BGJ01.q1**).
    g. Click on **Fit Q3**; when the fit statistics are displayed, click on **Save** on the **File** menu, choose the appropriate directory, and enter a file name (e.g., **BGJ01.Q3**).
    c. Estimation Summary: Click anywhere in the **Estimation Summary** section of the **Parameter Estimation** window, click on **Save (selected) Text** on the **File** menu, and choose the appropriate directory and name (e.g., **BGJ01.Sum**).
    d. Estimation Detail: Click anywhere in the **Estimation Detail** section of the **Parameter Estimation** window, click on **Save (selected) Text** on the **File** menu, and choose the appropriate directory and name (e.g., **BGJ01.dtl**).
    e. Save PARDUX window Contents: Click anywhere in the **PARDUX** window, click on **Save (selected) Text** on the **File** menu, and choose the appropriate directory and name (e.g., **BGJ01.prn**).

Evaluating the Calibration Output

The Research Monitors and Research Associates will review the calibration output and complete the Summary of Calibration form (Table A). In evaluating the results, the following questions should be answered.

1. Does the number of items in the calibration run match the number expected?
2. Does the number of cases at least meet the n-count specified in the design? If the case count falls short of the Design n, will this shortage jeopardize the soundness of the item parameter estimates?
3. Do the mean and standard deviation of the raw scores look reasonable?
4. Do the n-counts vary dramatically over items? If this occurs, while not expected, then the reason for discrepancies should be ascertained.
5. Do the item p-values look reasonable? Given the test and test administration characteristics noted above, the p-values are most likely to be above .30.

6. What number of items had a maximum A or default C in the PARDUX output? It is not typical for more than a few items to attain a maximum A, and no more than about 30% of the items might be expected to have default C's.
7. Consider the B-value range of the items. It is rare to see items with B-values below –3.00 or above +3.00.
8. With respect to the number of estimation cycles, it is unusual to see calibration runs that required 50 cycles for convergence. Any that do should be studied further. What item is noted to have had convergence problems? Was this item also flagged for poor fit?
9. Non-converging items for operational and anchor items also are rare. If you find more than, say, 3, some cause for this should be investigated.
10. When items are flagged for poor fit (where Z > (4*N)/1500)[1], their content should be checked to make sure that they have no unusual content properties. In addition, their observed and expected characteristic curves should be plotted using the plot routine within PARDUX. Capture and print the item.
11. Compare the anchor item parameters across the versions for each content area. Note any differences between the versions.

For pre-operational items that do not converge:

1. View the item summary and examine the item statistics
2. Run additional stages to see if convergence can occur with the current criterion.
3. Run an M-step on the individual item to see if it can converge without hand fitting.
4. Hand fit an item that did not converge in the M-step. Only "operational" items should be hand fit. **Field test items should not be hand fit.**

Once the initial calibration has been evaluated and hand fitting completed where required, the parameter file should be labeled appropriately and the parameters placed onto the Field Test Scale (see next section). NOTE:

- If all items converged and no hand-fitting was required, the prefix will be of form: CCM??.PAR
- If the M-step was used or if items were "turned off" after the first calibration, there will be a "B" in the prefix, e.g. CCM??B.PAR. A new folder should be constructed and the appropriate files placed into it (e.g. Version A would contain all files from the first calibration). If subsequent calibrations are required (i.e., after version B, additional items were turned off), the files would have a "C" in the prefix.
- If hand-fitting was required there will be an "H" in the prefix, e.g. CCM??H.PAR.

[1] See Wendy Yen's October 17, 1991 for a complete description of identifying potentially misfitting items.

Placement of Field Test Items onto the May 2000 Field Test Scale

The field test scale for the Maryland High School Assessment (MD HSA) will be defined by the May 2000 administration. All future field test items will be equated to this administration until the tests become fully operational and high-stakes are attached. The placement of the May 2001 field test items onto the Field Test Scale will be completed by common item equating using the predefined set of anchor items that were included on all of the 2000 forms. The Stocking-Lord (S-L) procedure will be used to align the anchor set test characteristic curves. Transformations will be required to place the parameters onto the Scale used for the creation of the DAT files and for pre-operational form construction (see section on Converting Parameters to the Field Test Scale). In addition, an additional transformation will be required once the final Scale is defined by MSDE. Note for item parameter estimation the C parameter should be "fixed" to the anchors.

PARDUX is used to complete the SL procedure and obtain the transformation values. To begin, transfer the following files to the folder

M:\PROJECTS\mryhsch\2001\May\FT\cc\equate:

- RWO, final CTL, and theta metric PAR files from the current administration
- ANC file containing the item parameters for the anchor items

Note the PAR and ANC files should be in the theta metric. In addition, the files should have the same root name e.g., CCM00.*

Where:
CC = content area
M = May administration
01 = year 2001 administration

The following steps outline the SL procedure using PARDUX:

Note for these analyses the C parameter should be "fixed".

Using Pardux

1. Read in the CTL file using the File/Read pull-down window. The RWO and ANC files are automatically read in if they have the same root name as the CTL file.

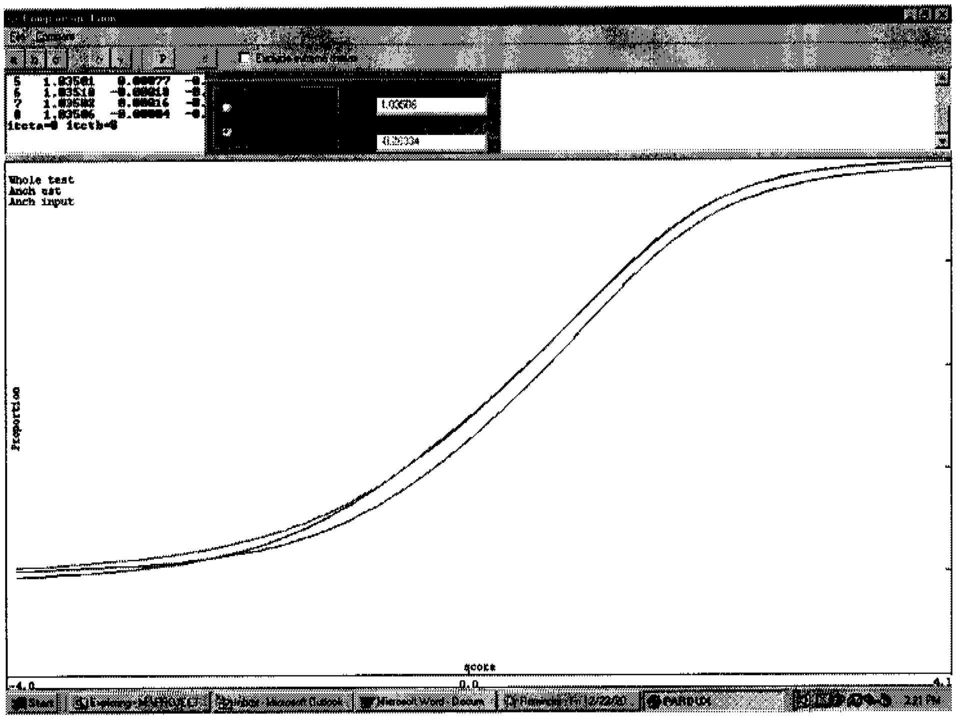
2. Read in the PAR file using the File/Read pull-down window.

3. Estimate thetas (Estimation/Thetas).

4. To complete the SL procedure, use the View/Comparisons. A split screen will appear with a graphics screen at the bottom and a text box at the top. A dialogue box also appears in the top part

of the screen (text). This dialogue box has options for transforming the parameters using the SL procedure. A single left click in the text part of the screen causes the dialogue box to toggle on and off. In the dialogue box under "Stocking and Lord", select "Yes (Transformed)" – this will provide the M^1 (multiplier) and M^2 (additive) values which will be used to calculate the m_1 and m_2 that will be used in the HLK files to transform the January parameters (see diagram on next page).

PARDUX Split Screen Compare Form Window



```
5   1.03501   0.00077  -0.
6   1.03510  -0.00010  -0.
7   1.03502   0.00016  -0.
8   1.03506  -0.00004  -0.
iteta=0 itetb=0
```

5. To compare the resulting transformation use Compare/Summary in the Compare form screen. Detailed information for each parameter is listed in the text area including, correlation, RMSD, the mean difference, the ratio of the standard deviation as noted below:

```
Maryland High School Assessment January Field Test Biology
       r     RMSD   mn diff  sd ratio   rdif   N  Lik = 1.0000000 E0
A  0.92   0.123  -0.017    0.844   -0.56   28    0 stages  BGJ01.PAR
B  0.98   0.202  -0.004    0.985   -0.18   28   M1 =   1.035 M2 = -0.263
C  0.81   0.040  -0.013    0.988   -0.32   28   adjusted
```

6. In addition compare and plot the a-, b-, c-, p-values, and TCCs using the Compare pull-down window ore the buttons under the menu.

7. Save the TCC and p-value plots by using the screen capture (alt+print scrn) and pasting the plots into a word document. Label each plot carefully.

8. Save the results of the transformation by using the File/Save (selected text) on the menu bar. Change the save as type to .sl and save the information in the content folder.

9. Close the Comparison Form window and save the data in the main Pardux screen by using the File/Save (selected text) on the menu bar. Change the save as type to .TXT, and use the extension .SLT and save the information in the content folder.

10. Save the transformed parameter estimates by using the File/Save Estimates on the menu bar include the letter "Z" (make sure file extension selected is *.par) after the root name to indicate that these are the final calibration parameters in the scale score metric after transforming using the SL procedure.

11. These are the files that we will use to construct DAT files. Summary information, final parameters, evaluation forms and plots should be placed in the following directory:

M:\PROJECTS\mryhsch\2001\May\FT\Cal\CC\DAT file Par

Evaluation

The attached Form B should be completed after the above process recording the M^1 and M^2 and information of the comparison between the two sets of p-value estimates (anchor and nonanchor). It is expected that between the two sets of p-value estimates the:

Standardized mean difference ≤ 0.05
Ratio of the standard deviation will be near 1.00
Correlation of the item parameters will be ≥ 0.90.

The plot of the TCCs provides a picture of the alignment that was achieved between the test based on the item parameters in the ANC file and the test based on the set of equated parameter estimates. Ideally these curves will line up so well that no differences between the input (ANC) and the estimated (May) TCCs are visible.

Note, comparisons between the a-, b-, and c- parameters before and after the equating should also be printed (see step 7 and 8 above).

After finalizing the transformation, save the transformed parameter estimates by using the File/Save Estimates on the menu bar include the letter "Z" after the root name to indicate that these are the final calibration parameters in the scale score metric.

M:\PROJECTS\mryhsch\2001\May\FT\equate\CC

In addition a copy of the final parameter files into the directory:

M:\PROJECTS\mryhsch\2001\May\FT\Datfiles\Final

Scoring Tables

Beginning January 2001, scores reported for the total test will be summary scores at the classroom, school and district levels. In 2002, reported scores will be at the student level and will appear on the student's transcript, although the score will not count toward high school graduation.

Because of the nature of the forms, items that contribute to the scores will need to be determined after examination of the classical and item response theory statistics. To accomplish this, the "best available" items in the pre-operational form will be selected to match the test blueprint. Content Editors will be provided with a list of items that were not calibrated (i.e., poor classical statistics, collapsed score level, did not converge) and an Excel spreadsheet with a column to indicated if the item contributes to the scores. Editors will code 0 if the item does not contribute and 1 if the item does contribute. Once completed this table will be used by Research in developing scoring tables and will be sent to Technology. The table that is sent to Technology must be a space delimited (*.prn) file and must contain all information for all forms and content areas.

Summary tables for schools, districts and the state will be provided based on the percent correct for those students that completed the pre-operational form.

Red Team Review

Prior to the release of the GRF to MSDE we will conduct analyses to independently verify:

- The items included in the calculation of student scores match the list supplied to technology (i.e., items contributing).
- The calculation of summary reports (i.e., percent of students within each score band, mean, standard deviation, min and max scores).

The information that will be supplied to MSDE includes summary reports of the percent of students within scoring categories for each School, LEA and the State. A relative frequency distribution for each score band will be supplied along with the mean, standard deviation, minimum and maximum scores for each content area. For January, Algebra will be excluded.

To complete this work, we will replicate the scores that will be assigned to each student and the summary tables using the GRF that will be provided to the customer. It is expected that there will be 100% agreement between the score reports and the summaries completed by Research. Any discrepancies must be identified and explained prior to the release of the Score Reports to MSDE.

Important considerations include:

1. Scores will only be reported for the "operational" items on the "pre-operational form. The item-contributing table sent to technology should be used to determine the items that contribute to student scores. Percent correct scores should be rounded to the first decimal place for each student and for all subsequent calculations.

2. Exclusion criteria for specific accommodations as outlined in the Technology Specifications must be duplicated exactly.

3. Students with missing LEA codes will be excluded from the summary tables. Students missing School codes will also be excluded from the summary tables.

4. Calculation of the percent correct will be based on the student raw score/total possible score points.
   - For Braille students, some items could not be Brailled in Algebra and Geometry. Therefore, the percent correct for these students should be calculated with a modified denominator.

5. Means and Standard Deviations of scores by School, LEA, and State should be rounded to the 1st decimal place. Min and Max Scores should not be rounded. The percent of students within each score band should be truncated at the 1st decimal place.

Preparation of DAT files

After all the WinPardux runs have been satisfactorily completed, Statistical Services will then up-load the files to the mainframe and produce .DAT files in ItemWin format. These .DAT files will be reviewed by Research prior to turning them over to Development for use in item selection. Note: Items excluded from the calibration or that didn't converge will not be included in the .DAT files. A list of excluded items is forwarded to the editors along with reasons for their omission from the pool. The editors then, using ItemWin, conduct interactive test selection for each content area.

The .DAT files will be posted in the following directories:

- M:\projects\mryhsch\2001\FT\Datfiles\CC, and
- the CTB-PUB network for Development's access for item selection.

Information about DIF should be included in column 42. A list of items flagged using Mantel and Mantel-Haenszel will be forwarded to Stat Analysis for inclusion in the DAT files. Each item will be flagged only once using the coding scheme listed below. If the item was flagged for more than one group (e.g. African American and female), only indicate the flag with the lowest number (e.g., African American).

1=African American
2= Hispanic
3=Asian
4=Native American
5=Female
6=Male

The standard .DAT file format that should be followed is listed below:

Line	Column	Information Shown
T line	1-1	“Content Area”
	14-15	Level
	17–60	DAT file name
L line	1-1	“LEVEL”
	7-7	Target level
	18-20	5th Percentile SS
	25-27	95th Percentile SS
P line	4-8	Sequential number (using the passage ID)
	10-29	Passage Descriptor (from the PEID)
O lines	1-1	“O”
	4 – 5	“OB”
	7 – 8	Objective number
	10-60	Objective title
I lines	1-1	“I” (for MC item) or “M” (for CR item)
	2 – 4	Booklet item number
	6-8	Content Level ID (i.e.,09, 10, 11, 12, 13, etc.)
	9-9	Anskey (mc options 1-5) blank for cr-items
	10-10	For cr-items: scrPnts+1; mc-items: “1”
	12 – 17	***a*** parameter
	19 – 26	***b*** parameter
	28 – 33	***c*** parameter
	35 – 36	Book number
	38 – 39	Objective number
	40-40	*- indicates anchor items
	42-42	Bias rating
	44-44	Fit rating
	46-50	PassageID
	52 – 71	Item descriptor: use Subskills Title
	72 – 79	ItemID (from the PEID)
	81--	***g3, g4,*** etc. For CR item

NOTE:

- Numbers should be right-justified (leading blanks or zeros)
- Letters should be left-justified (trailing blanks). If “P” line has only 000 Passage number, Don’t show the “P” line.
- In the “I” line, the columns 72-79 (PEID ITEM_ID) have to be unique for each item.

TABLE A
Summary of Calibration Results

Content Area	No of Items	Sample Size	Raw Score		P-Values	Max A	Default C	B-Value Range	N. of Est. Cycles	Non-Conv Items	Items Flagged for Poor Fit
			Mean	SD							

Table B
Maryland High School Assessment
Evaluation of Equated Item Parameters

Research Monitor: ______________________

Date Completed: ________________

Test	N of Anchors	P-value Comparison After Equating				M1	M2
		Diff	RMSD	SD Ratio	r		
Algebra							
Biology							
English							
Geometry							
Government							