November 20, 2003

Summary of Methods and Procedures for Calibrating, Equating and Scoring the 2003 Maryland High School Assessments

I. Calibrating and Equating the January 2003 Test Forms

Algebra, Biology, English, and Government

Our original calibration and equating design for the Maryland High School assessments required within-classroom spiraling of the new test forms with Form W, which was a repeated test from 2002 (for Government, this corresponded to the May 2002 Form V; for all other content areas, this was the May 2002 Form W). However, the spiraling of test books did not produce the desired randomly equivalent groups across all forms, and forms that occurred early in the spiral were administered to a disproportionately high number of ESL and special education students. We also found that the test characteristic curves (TCC's) for Form W (which contained no items in common with the other forms) did not always parallel the TCC's for the new test forms. The difference between TCC's was minimal for Algebra and fairly small for Biology and Government, but was very large for the English test.

Because we did not have randomly equivalent groups for all forms, and because Form W had no items in common with the other forms, we adopted the following calibration and equating strategy:

1. Calibrate all items across test forms, using the May 2002 scale-score item parameters for multiple-choice items in Form W as anchors in a Stocking and Lord procedure.

2. Examine the results for reasonableness.

    a. Compare the TCC's and scale score distributions for Form W in May 2002 and January 2003.

        - Results for each content area indicated that Form W behaved similarly in the two administrations (i.e., similar means and standard deviations, similar TCC's).

    b. Score the repeated Form W using the old and new item parameters (from May 2002 and January 2003) and compare the results.

- Results for each content area indicated that the two sets of parameters produced similar results, with very few substantial score discrepancies between the two.

c. Compare test scores and demographic characteristics by test form.

- Results for each content area indicated that demographic characteristics of the students completing Form W were very similar to the characteristics of students completing the last new form in the spiral (i.e., English Form C, Biology Form C, Government Form D, and Algebra Form B).

- As noted above, the results indicated that the first form(s) in each spiral had a disproportionately high representation of ESL and Special Ed. students, and substantially lower test scores than the other new forms.

3. Because the demographic characteristics of examinees taking Form W were found to be comparable to the characteristics of students taking the last of the new forms, item parameters for the last new form in the spiral were equated to Form W. For this transformation, the original (May 2002) item parameters were used to score Form W. CTB's WinFlux software was then used to find the linear transformation that best approximated equipercentile score equating.

4. The transformation constants from this procedure were then applied to the item parameters for all of the new test forms.

5. The transformed parameters were used to score all new forms. Based upon the results of Step 2b, Form W was scored using the original May 2002 parameters. Raw score and scale score distributions were generated for all test forms, and these distributions were reviewed for reasonableness (i.e., results were compared across forms, and were compared with results from the January 2002 and May 2002 administrations).

Geometry

Unfortunately, Maryland experienced a severe snowstorm on the date scheduled for the Geometry test administration, and fewer than 3,000 students completed the test on that date. As a result, the spiraled sample contains very few students (approximately 700 usable cases per form) from a very small number of LEA's. The vast majority of students in the spiraled sample come from Anne Arundel, Frederick, Howard, Washington, and Baltimore City.

More than 6,000 students were tested at a later date with a makeup form. Approximately 4,900 received makeup form X, and approximately 1,200 received makeup form Y.

In light of this situation, we modified our original calibration and equating plan as follows:

1. First, calibrate only the 22 core items for all new test forms, including make-up forms. By calibrating the same items for all students, we eliminate the need for randomly equivalent groups and are able to include both the primary form examinees and the makeup-form examinees in our initial calibration sample.

2. Calibrate the remaining items, fixing the parameters for the 22 core items at the values obtained in Step 1, above.

3. Use the May 2002 scale-score item parameters for multiple-choice items in Form W as anchors in a Stocking and Lord procedure.

    a. Examine the results for reasonableness.

    b. Compare the TCC's and scale score distributions for Form W in May 2002 and January 2003.

    c. Score the January 2003 Form W using the old and new item parameters and compare the results.

        - Results indicated that the two sets of parameters produced similar results, with very few substantial score discrepancies between the two.

    d. Compare scores and demographic characteristics by test form.

        - Results indicated that demographic characteristics of the students completing forms C and W were very similar to one another.

        - However, Forms A and B included disproportionately high representation of ESL and Special Ed. students.

        - Results indicated that Form W was behaving similarly in the two administrations (i.e., similar means and standard deviations, similar TCC's).

        - However, the TCC for Form W was somewhat different from the TCC's for the other forms, and the mean scale score on Form W

was substantially different from the mean scale scores on the new test forms.

4. Because the demographic characteristics of examinees taking Form C and Form W were comparable, item parameters on Form C were equated to Form W. For this transformation, the original (May 2002) item parameters were used to score Form W. CTB's proprietary WinFlux software was then used to find the linear transformation that best approximated equipercentile score equating.

5. The transformation constants from this procedure were then applied to the item parameters for all of the new test forms.

6. The transformed parameters were used to score all new forms. Based upon the results of Step 3c, Form W was scored using the original May 2002 parameters. Raw score and scale score distributions were generated for all test forms, and these distributions were reviewed for reasonableness (i.e., results were compared across forms, and were compared with results from the January 2002 and May 2002 administrations).

Tables 1a and 1b show the transformation constants from the Stocking & Lord and Winflux equating procedures, and the resulting score means and standard deviations by test form.

Table 1a. January 2003 Transformation Constants Equating New Forms Back to Repeated Test Form (W) from May 2002*

Content Area	Stocking & Lord		Linear Equipercentile*	
	M1	M2	K1	K2
Algebra	31.957	406.717	1.030	-12.563
English	33.104	393.867	0.800	86.643
Biology	35.058	400.754	0.999	4.160
Geometry	30.690	407.376	1.060	-31.590
Government	35.136	499.795	0.982	9.731

* Within each content area, the same K1 and K2 values were applied to the Stocking & Lord equated item parameters in all of the new forms (i.e., all forms except Form W, which was scored using the original May 2002 parameters).

Table 1b. Summary of Maryland HS Equating Results
Research Sample, January 2003 Administration
Equating Back to Repeated Test Form (W) from May 2002*

Content Area	Test Form		After Stocking & Lord		After Linear Equipercentile	
	Form ID	N of Cases	Mean	S.D.	Mean**	S.D.
Algebra	A	2027	399.00	42.71	398.52	43.60
	B	1787	402.84	40.20	402.44	41.06
	W*	1762	402.30	41.42	--	--
English	A	2370	380.75	45.80	390.84	38.07
	B	2090	387.60	41.74	396.42	34.61
	C	2019	386.52	41.62	395.54	34.61
	W*	1986	395.46	34.43	--	--
Biology	A	2239	392.10	41.17	394.83	41.28
	B	1938	395.23	40.76	397.94	40.82
	C	1890	396.23	40.81	398.97	40.85
	D	1819	397.23	39.23	399.97	39.29
	W*	1819	398.79	40.78	--	--
Geometry	A	5610*	405.10	39.54	397.74	39.97
	B	1880*	397.59	37.59	389.52	38.61
	C	640	406.03	39.15	398.43	40.46
	W*	629	394.08	41.72	--	--
Government	A	2380	393.75	41.03	395.08	40.48
	B	2038	397.12	39.95	398.39	39.38
	C	2012	397.49	38.45	398.79	37.85
	D	1956	398.05	38.98	399.31	38.34
	W*	1917	399.15	38.34	--	--

* Repeated test forms were scored with their May 2002 item parameters.

** Note that because an equipercentile approximation was used, the means and standard deviations for Form W do not necessarily match the equated means and standard deviations for the last operational forms in the spiral. Geometry is the most discrepant case because of differently skewed distributions.

*** For Geometry, Form A includes 740 spiraled forms and 4,870 non-spiraled makeup forms; Form B includes 661 spiraled forms and 1,219 non-spiraled makeup forms.

II. Calibrating and Equating the May 2003 Test Forms

In May 2003, all new tests with form designations of L or below shared the same core of common items that were included in all of the January 2003 test forms. Because of this strong core, we were able to equate these new forms and place them on the January 2003 scale by calibrating all forms together and using the core items as anchors in a Stocking and Lord equating procedure.

However, this approach was not appropriate for scaling and equating the block field test forms (Forms M and above), because those forms shared no common items with each other or with the primary operational test forms. Therefore, we adopted the following strategy:

1. Within each content area, calibrate all items (from all forms) together, and use the common items from January as equating anchors in a Stocking and Lord procedure.

2. Using a linear approximation to equipercentile equating in WinFlux, equate the item parameters in each block field test form (Forms M and above) to the last operational test form in the spiral (Form L). Obtain a set of unique transformation constants for each block field test form.

3. Apply the transformation constants to Forms M and above to put them on scale.

4. Produce raw score and scale score distributions for all test forms, and review these distributions for reasonableness. Compare results across forms and across administrations.

Tables 2a and 2b show the transformation constants from the Stocking & Lord and Winflux equating procedures, and the resulting score means and standard deviations by test form.

Table 2a. May 2003 Transformation Constants
Equating New Forms to January 2003*

Content Area	Stocking & Lord		Linear Equipercentile*		
	M1	M2	Form ID	K1	K2
Algebra	39.034	416.066	M	0.965	18.329
			N	0.959	18.617
English	31.567	399.434	M	0.930	37.772
			N	0.991	9.938
			P	0.935	29.298
Biology	39.231	407.390	M	0.938	29.051
			N	0.932	30.296
			P	0.943	26.907
			Q	0.980	11.315
			R	0.951	22.762
Geometry	32.495	406.495	M	0.984	10.673
Government	39.726	405.821	M	1.035	-6.394
			N	1.049	-14.006
			P	0.994	15.883
			Q	1.034	-3.749
			R	1.004	7.946
			S	1.032	-4.015

Linear approximation to equipercentile equating was applied only to the block field test forms (i.e., forms M and above), equating each form individually to Form L. The Stocking and Lord transformations were applied to all forms prior to the linear equipercentile.

Table 2b. Equating Results for Maryland HS Assessments
Research Sample, May 2003 Administration
Equating Back to January 2003

Content Area	Test Form		After Stocking & Lord		After Linear Equipercentile*	
	Form ID	N of Cases	Mean	S.D.	Mean	S.D.
Algebra	C	6585	405.8	51.4	--	--
	D	5639	413.9	46.0	--	--
	E	5559	414.5	45.9	--	--
	F	5468	414.6	45.0	--	--
	G	5420	412.6	45.9	--	--
	H	5405	413.4	45.3	--	--
	J	5340	414.0	45.5	--	--
	K	5256	414.2	45.6	--	--
	L	5194	413.9	45.0	--	--
	M	5134	410.2	46.0	413.9	44.7
	N	5069	412.2	46.5	414.0	45.0
English	D	5831	390.3	39.8	--	--
	E	4797	397.7	34.5	--	--
	F	4806	398.0	34.8	--	--
	G	4772	397.1	34.3	--	--
	H	4775	397.5	35.5	--	--
	J	4720	397.9	35.5	--	--
	K	4673	399.4	34.4	--	--
	L	4600	398.8	36.2	--	--
	M	4596	388.1	38.7	398.7	36.7
	N	4508	393.0	36.9	398.7	36.7
	P	4483	395.3	37.7	398.9	35.9
Biology	E	4965	399.8	47.0	--	--
	F	4147	405.8	43.6	--	--
	G	4126	406.6	42.9	--	--
	H	4099	406.5	42.1	--	--
	J	4123	406.4	43.6	--	--
	K	4051	405.0	43.9	--	--
	L	4017	405.6	42.3	--	--
	M	4004	401.4	45.0	405.5	42.7
	N	3970	402.8	45.3	405.2	43.3
	P	3910	401.5	44.9	405.4	43.1
	Q	3866	402.2	42.9	405.6	42.3
	R	3821	402.6	44.4	405.5	42.9

* Note that because an equipercentile approximation was used, the means and standard deviations for Form L do not necessarily match the equated means and standard deviations for the block field test forms.

Table 2b. Equating Results for Maryland HS Assessments
Research Sample, May 2003 Administration
Equating Back to January 2003
(Continued)

Content Area	Test Form		After Stocking & Lord		After Linear Equipercentile	
	Form ID	N of Cases	Mean	S.D.	Mean	S.D.
Geometry	D	5790	399.3	43.5	--	--
	E	5069	402.6	42.6	--	--
	F	4993	404.3	40.0	--	--
	G	4997	404.1	41.2	--	--
	H	4912	403.6	42.3	--	--
	J	4874	404.6	39.7	--	--
	K	4813	403.8	41.4	--	--
	L	4782	404.5	39.1	--	--
	M	4668	400.3	40.0	404.4	40.0
Government	E	4897	398.7	49.5	--	--
	F	4005	407.9	45.5	--	--
	G	4037	408.4	46.0	--	--
	H	3957	407.8	44.6	--	--
	J	3971	408.8	46.3	--	--
	K	3954	409.0	45.5	--	--
	L	3977	409.1	45.5	--	--
	M	3900	401.6	43.9	409.2	45.6
	N	3849	403.3	43.5	409.1	45.5
	P	3805	396.3	45.7	408.6	45.9
	Q	3773	400.1	43.9	409.9	45.5
	R	3726	400.7	45.2	410.2	46.0
	S	3644	401.1	44.0	409.8	45.6

III. Scoring the January and May 2003 Test Forms

All tests were pattern-scored using the final equated IRT item parameters. Because pattern scoring was used, students obtaining the same raw score on a particular test form usually do not receive the same scale score. Nevertheless, tables of estimated raw-to-scale score values and standard errors were produced for each test form, and were used as initial values in the IRT scoring algorithm. These tables will be delivered to MSDE under separate cover.