

NAEP-QA FY06 Special Study: 12th Grade Math Trend Estimates

Tirso E. Diaz
Huy A. Le
Laurens L. Wise

Prepared for: U.S. Department of Education
National Center for Education Statistics
600 Independence Ave.
Washington, D.C. 20202

Prepared under:

MOBIS Contract No: GS-10F-0087J
Order No.: ED-02-PO-3120
Delivery Order: 0019

December 19, 2006

Opinions contained in this document reflect the views of the authors and do not reflect the views of the US Department of Education or National Center for Education Statistics.

NAEP-QA FY06 Special Study: 12th Grade Math Trend Estimates

Tirso E. Diaz
Huy A. Le
Laurens L. Wise

Prepared for: U.S. Department of Education
National Center for Education Statistics
600 Independence Ave.
Washington, D.C. 20202

Prepared under:

MOBIS Contract No: GS-10F-0087J
Order No.: ED-02-PO-3120
Delivery Order: 0019

December 19, 2006

Opinions contained in this document reflect the views of the authors and do not reflect the views of the US Department of Education or National Center for Education Statistics.

NAEP-QA FY06 Special Study:
12th Grade Math Trend Estimates
Table of Contents

Overview and Purpose of this Report 1

Methods..... 1

 Available Data 1

 Differences Across Administrations 2

 Estimating Trends 4

Results 8

Discussion..... 9

 Limitations 9

List of Tables

Table 1. Common Items in Each Block for the 2000 and 2005 Administrations 3

Table 2. Mean Differences in Proportion Correct Statistics for Common Items..... 4

Table 3. Frequencies of Students Responding to Link Items in 2000 and 2005 6

Table 4. Key Statistics for Proportion-Correct Scores Based on Common Items 7

Table 5. Statistics of Link Items Used in Estimation of Coefficient Alpha 7

Table 6. Posterior Theta Distribution Estimates from Joint Calibration 8

Table 7. Estimated Score Gains Using Each Method 9

Table 8. Increase in Proportion Correct by Item Position in 2005 10

List of Figures

Figure 1. Comparison of Proportion Correct Statistics for Common Items in the 2000 and 2005 Grade 12 Math Assessments 2

Overview and Purpose of this Report

The National Assessment Governing Board (NAGB) revised the prior mathematics content frameworks for the 2005 NAEP assessment. Changes to the frameworks for the grade 4 and 8 assessments were judged to be minor and so normal NAEP equating procedures were used to produce estimates of trends in mathematics achievement for these grades. Trend reports for each state and for the nation were reported in Fall 2005.

Results from the 12th grade, national-only assessment are scheduled to be released in February 2007. Framework changes for Grade 12 were judged to be more major and current plans for this release do not include any information on achievement trends. A bridge study to estimate trends under the old content framework was proposed by ETS but not funded.

Particularly with increased attention to student achievement at grade 12, the absence of information on trends in mathematics achievement at this level is unfortunate. While a formal bridge study was not implemented, there was nonetheless considerable overlap in the test questions used with the 2005 grade 12 mathematics assessment and the prior grade 12 mathematics assessment in 2000. Specifically, 118 test questions were used in both assessments, albeit in different block and booklet positions and structures.

The present study sought to investigate the available information on trends in mathematics achievement at grade 12 from 2000 to 2005. This report summarizes the main procedures used and findings from this study. It was undertaken as part of the NAEP Quality Assurance (NAEP-QA) contract as a study of potential linking processes for tracking trends across future changes in NAEP content frameworks.

Methods

Two divergent methods were used to estimate differences in grade 12 mathematics achievement from 2000 to 2005 in standard deviation units. These estimates were then converted to gain scores in the NAEP 2000 reporting scale metric. The gain scores for all students and for gender and race subgroups were tested for difference from zero. This section describes the data and methods used in developing and testing the estimated gain scores.

Available Data

ETS provided an Excel spreadsheet with data on 118 test questions (items) included in both the 2000 and 2005 grade 12 mathematics assessments. The spreadsheet showed block and items numbers for each administration, along with weighted estimates of the proportion passing (p-value) the item (or mean score for polytomously scored items) and the proportion of students omitting or not reaching the item. Figure 1 provides a plot of the corresponding p-values from the 2000 and 2005 administrations.

**NAEP Math Grade 12 - Common Item Proportion Correct Statistics:
2005 (25-min blocks) vs 2000 (15-min blocks)**

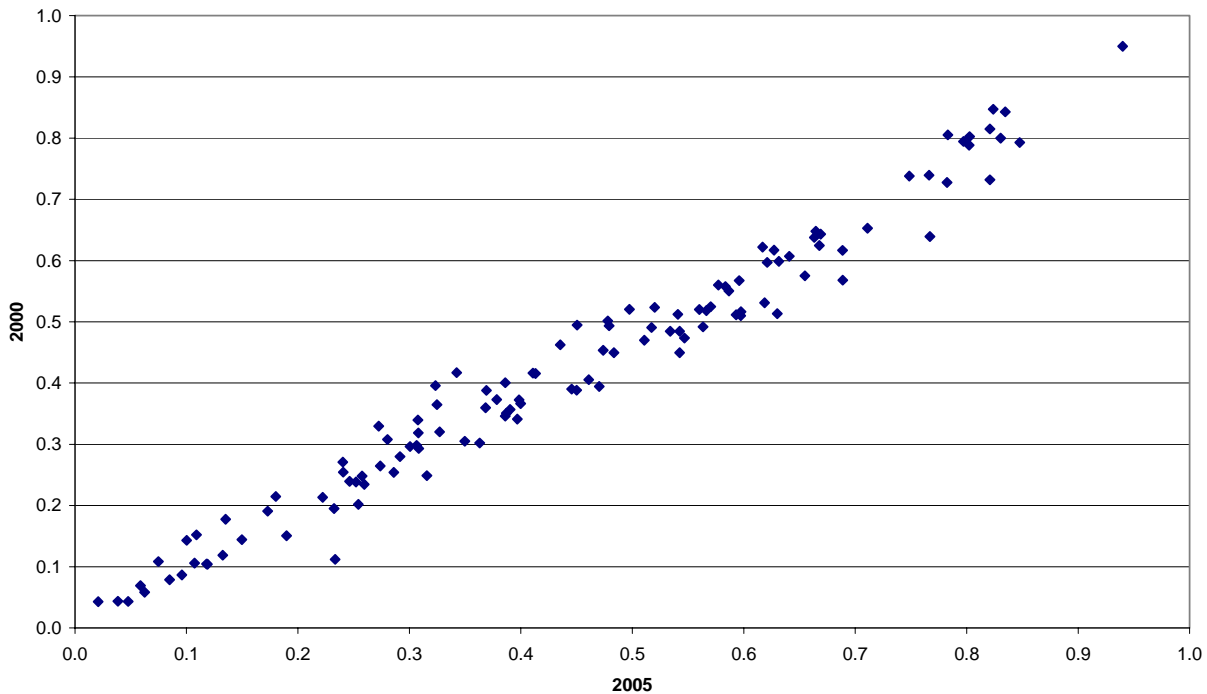


Figure 1. Comparison of Proportion Correct Statistics for Common Items in the 2000 and 2005 Grade 12 Math Assessments.

ETS also provided analysis files containing student records from each of the two administrations. The analysis files included item responses and scores for all of the test items, including the 118 common items, demographic information and a variety of sampling weight and sample identification variables. File documentation and codebooks were also provided for each of the two analysis files.

Differences Across Administrations

Booklet Structure. There were several important differences between the 2000 and 2005 administration that might affect trend estimates generated from the common items. First, the booklet structure was changed from three 15-minute blocks in 2000 to two 25-minute blocks in 2005. As a consequence, the position of individual items within blocks and booklets was different between the two assessments, although position shifts were balanced across the set of items. Table 1 shows the blocks used with each administrations and the number of common items included within each block.

Table 1. Common Items in Each Block for the 2000 and 2005 Administrations¹

2000 Grade 12 Math Assessment		2005 Grade 12 Math Assessment	
Block	Common Items	Block	Common Items
C	13	C	17
D	3	D	17
E	8	E	18
F	10	F	4
G	8	G	5
H	11	H	21
I	6	I	18
J	9	J	18
K	14	K	0
L	9	L	0
M	9		
N	9		
O	9		
Total	118	Total	118

Accommodations. The 2000 assessment included a bridge study of the impact of allowing increased accommodations. For half of the schools, students were allowed only limited accommodations, consistent with prior assessments. For the other half of the schools students were allowed a broader set of test accommodations. The availability of additional accommodations increased inclusion rates among students with disabilities (SD) and English Language Learners (ELL). Results based on SD and ELL students from the no-accommodation sample (and all other students) were reported in 2000 to provide appropriate comparisons with prior assessments. In the 2005 administration, the additional accommodations were available to all SD. Results in the present study are based on the SD and ELL students in the second (accommodated) sample of 2000 students.

Calculator Use. Policies regarding the use of calculators were changed between the 2000 and 2005 administrations. These changes might have affected results for some items requiring

¹ Each 2000 booklet contained three blocks of test questions, while each 2005 booklet contained 2 blocks.

computations and not other, non-computational items. The plot shown in Figure 1 indicates a high degree of consistency across all items in the relationship of 2000 and 2005 results. If changes in calculator use policy had a significant impact, we would expect to see differences in trend lines between calculator and non-calculator items. Given the lack of such differences, the issue was not further addressed in this study.

Scoring of Constructed Response Items. A few of the constructed response items were treated differently in operational analyses of the 2000 and 2005 test results. Based on preliminary analyses, ETS makes decisions about whether and how to collapse score levels to eliminate low-frequency levels and increase overall stability. Where different decisions about score levels were reached in 2000 and 2005, we applied the 2005 score level treatment to the 2000 data for consistency.

Estimating Trends

Table 2 shows a simple analysis of the item-level results provided by ETS. The analysis indicates a statistically significant increase in the percent of students answering each question correctly (from 42% to 44%). Similar results are shown for a logistic transformation of the proportion correct statistics. These results do not, however, translate directly to gains on the reporting scale which is based on student-level estimates rather than item-level estimates. Two divergent methods were used to assess student-level gains. The first approach, the proportion correct metric analyses, corresponds closely to the item level results. The second approach, involving a joint item response theory (IRT) calibration, corresponds more closely to the process used to link results.

Table 2. Mean Differences in Proportion Correct Statistics for Common Items

Statistic	Proportion Correct	Logits ¹
2000 Mean	.4212	-.1771
2005 Mean	.4418	-.1409
Average Difference	.0206	.0361
Standard Deviation of Differences	.0404	.0963
t-statistic (based on 118 items)	5.51	4.06

¹ Logits, defined as $\log(p/(1-p))$, are often used to convert proportions to a more normalized metric.

Proportion Correct (PC) Scores

For each student in the 2000 and 2005 assessment, we computed the proportion of the common items answered correctly. Scores for items with more than two score levels were scaled to reflect the proportion of the maximum possible score obtained by each student. Within each test block, items that the student did not reach were treated as “not presented” and excluded from the proportion-correct calculations. Differences in the number of common items included in

different test booklets and in the number of these items reached by each student, led to variation across students in the number of items used in computing the proportion correct scores. Table 3 below shows the distributions of numbers of items reached by students in the 2000 and 2005 assessments.

As can be seen in Table 3, there is only a small number of cases where students responded to fewer than 4 common items, so these cases were eliminated from further analyses. Table 4 shows key statistics for the proportion-correct scores in the 2000 and 2005 samples.

Table 3. Frequencies of Students Responding to Link Items in 2000 and 2005

Number of Items Taken	2000 Sample	2005 Sample
	%	%
0	0.61	2.16
1	0.01	0.05
2	0.01	0.55
3	0.02	0.70
4	0.01	5.75
5	0.03	5.11
6	0.03	0.10
7	0.00	0.10
8	0.06	0.14
9	0.15	1.79
10	0.06	0.18
11	0.09	0.27
12	0.13	0.30
13	0.21	0.28
14	0.44	0.73
15	0.51	1.10
16	0.86	1.32
17	0.92	7.74
18	4.07	10.97
19	1.81	1.37
20	1.49	1.43
21	7.68	6.63
22	4.71	8.11
23	5.17	4.89
24	10.16	0.41
25	5.04	1.40
26	10.25	1.75
27	6.75	0.29
28	4.44	0.36
29	5.23	0.67
30	6.98	0.64
31	7.79	1.13
32	8.45	1.18
33	5.85	1.90
34	0.00	2.54
35	0.00	11.01
36	0.00	6.79
37	0.00	0.48
38	0.00	3.71
39	0.00	3.98
Total	100.00	100.00
Approximate Number of Students	13,700	9,300

Table 4. Key Statistics for Proportion-Correct Scores Based on Common Items

Sample	N (Wtd.) ¹	Mean	Std. Dev.	Reliability (Coefficient Alpha) ²	Correlation	
					with Plausible Values	Correlation Corrected for Error ³
2000 Sample ⁴	13,600	0.411	0.204	.865	.881	.985
2005 Sample	9,000	0.432	0.220	.844	.826	.939

¹ Rounded to nearest 100.

² Coefficient Alphas were estimated based on (a) average number of items the students took and (b) average inter-item correlations.

³ These correlations were disattenuated from the observed correlations and coefficient alphas in column 4 and average correlations among the plausible values (theta – range from .916 to .925).

⁴ This includes accommodated SD/LEP students in sub-sample 3 and weighted by AWEIGHT, as specified by ETS.

NOTE: Estimates are based on cases with at least 4 non-missing responses on the link items.

The difference in mean proportion correct scores for students, .02, is the same as the difference in proportion correct scores for items reported earlier. Deleting students who did not respond to a minimum number of items did not make an appreciable difference. In addition to score means and standard deviations, Table 4 provides an estimate of the reliability of the percent-correct scores and the correlation of these scores with the plausible values used in operational reporting. Information on which estimation of reliabilities for the proportion-correct scores was based is provided in Table 5. It can be seen that the inter-correlations among the link items appear to be very similar in the 2000 and 2005 samples.

Correlations corrected for measurement error (in both the proportion correct scores and the plausible values) indicate the relationship of the constructs measured by the proportion correct scores and the plausible values. While the corrected correlations are high for both samples, they are particularly high for the 2000 sample. Thus the common items provide a reasonable basis for assessing trends that would have been observed had the content framework used with the 2000 assessment been repeated in 2005.

Table 5. Statistics of Link Items Used in Estimation of Coefficient Alpha

	Number of link items taken				Inter-item correlation ¹			
	Mean	SD	Min	Max	Mean	SD	Min	Max
2005 Sample	23.49	10.44	4	39	.190	.115	-.23	.67
2000 Sample	25.98	4.73	4	33	.198	.109	-.21	.64
All sample	24.98	7.64	4	39	.195	.100	-.21	.61

¹ These are statistics across 6,903 pairs ($k(k-1)/2$ with $k=118$) of inter-item correlations.

NOTE: The statistics are based on cases with at least 4 non-missing responses on the link items.

Joint Calibration (JC)

The second approach involved a joint analysis of item responses in both the 2000 and 2005 administrations. A calibration was run using all of the items administered in each of the two assessments, not just the common items, so that results would correspond as closely as possible to operational results based on the full set of items in each year. The commercially available version of Parscale was used in these analyses. We had previously used the same software to generate item parameter estimates for each year separately and found a close correspondence with the parameter estimates used operationally by ETS.

In normal NAEP operations, item parameter estimates from a joint calibration like the one performed here would be used along with conditioning to provide posterior distribution estimates (theta scale) for each student from which plausible values are drawn. In the present study, we were primarily interested in estimates of the mean and standard deviation of the marginal distributions for each of the two populations. These estimates are shown in Table 6. The estimates were used to create a linear adjustment to the 2005 plausible values so that their mean and standard deviation would have the same relationship to the plausible values from the 2000 administration. Targets for the standard deviation and mean of the 2005 plausible values were computed as:

$$SD_{2005rep} = SD_{2000rep} * (SD_{2005theta} / SD_{2000theta})$$

$$MN_{2005rep} = MN_{2000rep} + (MN_{2005theta} - MN_{2000theta}) * (SD_{2000rep} / SD_{2000theta})$$

Table 6. Posterior Theta Distribution Estimates from Joint Calibration

Posterior Distribution Statistic	Population	
	2000	2005
Mean	-.0525	.0408
S.D.	1.0088	.9812

Results

Table 7 shows the estimated score gains using each of the two approaches. Estimated t-statistics indicating the significance of the gains are also shown. The standard errors used in computing these t-statistics were computed using jack-knife weights to account for sampling variation as well as estimates of measurement error. The two methods led to very similar estimates of score gains and also similar estimates of the statistical significance of these gains, not only for all students, but also for each gender and race group. The only difference in statistical significance was for Hispanics where the t-statistic from the proportion-correct score

gains was just slightly below the critical cutoff (1.94 compared to 1.96) while the corresponding statistic from the joint calibration method was slightly above the cutoff (2.18 compared to 1.96).

There was some variation across the two methods in the estimation of subgroup gains. For the PC method, estimated effect sizes were corrected (upward) to account for measurement error. The impact of this correction was small. The JC method used the plausible values developed by ETS based on conditioning on background variables.

Table 7. Estimated Score Gains Using Each Method

Group	2000 Mean	Using PC Scores		Using Joint Calibration	
		Score Gain	t statistic	Score Gain	t statistic
Total	300.2	3.4	3.44	3.3	2.83
Male	301.9	2.4	2.01	2.8	1.97
Female	298.6	4.4	3.95	3.7	3.05
White	307.4	4.3	3.78	3.7	3.02
Black	272.9	7.8	3.48	6.5	2.90
Hispanic	281.4	4.7	1.94	5.1	2.18
Asian	317.2	0.8	0.17	-0.8	-0.22

Note. Bold indicates statistically significant gains.

Discussion

The two approaches to assessing gains in 12th grade mathematics achievement provided roughly similar results. At the overall level, differences were scarcely larger than rounding. The somewhat larger differences in estimates for different subgroups may warrant further investigation. The results from the divergent procedures do provide potentially useful information on trends in Grade 12 mathematics achievement.

Limitations

As noted above, there were several differences between the 2000 and 2005 administrations that could have contributed to the apparent gain in test scores. Three main differences are discussed further here: (1) changes in booklet structure, (2) changes in calculator use policy, and (3) changes in test content.

Changes in Booklet Structure

Booklets in the 2000 administration contained three separately timed blocks, while the 2005 booklets contained two separately timed blocks. Each of the 2005 blocks was roughly fifty percent longer than the blocks used in the 2000 administration, increasing possible fatigue effects and possibly lowering performance on the 2005 blocks. Prior ETS studies indicated that scores were actually lower for the reconfigured booklets at grades 4 and 8 (unpublished ETS study, personal communication). No formal study was done of the effect of the change in booklet structure for 12th grade mathematics.

Table 8 shows a comparison of increases in item passing rates for the 78 common items in the first 13 positions of the 2005 test blocks and the 40 common items in positions 14 and later. Items in the first two-thirds of each 2005 block (roughly positions 1 through 13) would have been, on average, in similar positions in the 2000 blocks. Of course, some items would move forward and others would move back affecting omit and not reached rates as well as possibly affecting the proportion passing. On average, however, the effects of forward and backward movement in item position would tend to cancel out. Items in the later positions of the 2005 blocks would, necessarily, have been in earlier positions in the 2000 blocks, since these blocks were shorter. Thus there would be a systematic change to later item positions for these items and a possible depression in item performance due to fatigue (or, perhaps more likely for 12th graders, just a drop-off in motivation to continue).

Table 8. Increase in Proportion Correct by Item Position in 2005

Item Position	Using PC Scores		
	Mean Difference	S.D. of Differences	t statistic
All common items	.0206	.0404	5.51
1-13 (First two-thirds)	.0289	.0388	6.58
14+ (last third)	.0043	.0389	0.69
Difference (Early – Late)		Std. Error	
	.0247	.0076	3.27

Items in the first two-thirds of the 2005 block positions (no net change in item position) did show significantly larger increases in proportion correct compared to items toward the end of the 2005 blocks. The estimated increase using only items in the first two-thirds of the 2005 test booklets was 41% more than the increase estimated using all of the common items. This would translate into an estimated scale score gain of 4.8 points rather than the overall estimated gain of 3.3 points shown for the proportion-correct method in Table 7. While these analyses do not constitute a carefully planned study of booklet design effects, our best estimate is that the apparent gain of 3.3 scale score points may be quite conservative and could have been as much

as 1.5 points higher if the booklet design had not been changed. These results are consistent with ETS findings for grades 4 and 8 using a more carefully designed study.

Changes in Calculator Policy

If changes in calculator policy had a significant effect on the increase in performance on the common items, the increase would not be expected to be uniform across all items. Many items involve little or no computation and so would not be affected by the use of calculators. While we were not able to identify specific items that might have been affected, the consistency in increases across all items, as shown in Figure 1 above, suggests that changes in calculator use policy had a minimal effect, if any, on the increases calculated here.

Changes in Test Content

The common items analyzed here were a large, but not necessarily representative sample of all of the items used in the 2000 and 2005 assessments. The high correlations shown in Table 4 above between the proportion-correct scores based on these items and the plausible values estimated from all items in each assessment suggest that the common items do represent the content of the overall assessments reasonably well. This assertion is particularly true for the 2000 assessment, where the correlation between the common-item and overall scores was .985 after correcting for measurement error. The correlation was somewhat less, .939, for the 2005 assessment. Thus estimates on increases in achievement more closely reflect how students gained on the content of the 2000 assessment than on the 2005 assessment which included some new content.

Conclusions

The analyses described above indicate probable gains in 12th grade mathematics between 2000 and 2005. Because of the limitations noted above, it is not possible to completely rule out factors other than increased student performance that may have contributed to these gains. In the future, carefully designed bridge studies would enable much clearer estimates of gains in student achievement during the transition between content frameworks.