DRAFT: not for circulation except to addressees

June 20, 2010

MEMORANDUM

To: David Steiner

cc: John King, David Abrams, Alan Ray, Howard Everson

From: Dan Koretz

Subject: **Evidence about the leniency of 8th-grade standards**

At the April 22 TAG meeting, you requested three types of evidence bearing on the leniency or severity of the performance standards on the New York State tests:

1. The relationship between achieving the cut scores on the 8th-grade tests and later Regents exam scores;
2. The relationship between the 8th-grade cut scores and performance on NAEP; and
3. The relationship between Regents exam performance and later performance in the CUNY system.

In the subsequent discussion, the TAG added a fourth type of evidence:

4. Impact data, that is, the effects of raising the cut scores on pass rates.

This memo provides findings with respect to the first two of these four questions. Howard Everson will separately send a memo addressing the third. Because we agreed that the fourth question should be addressed using data from the 2010 assessment, which only CTB has at this time, we agreed that Rich Patz would address the fourth question.

The comparison with NAEP shows that some of the New York performance standards are lenient and that they have been becoming rapidly more so. The comparison between 8th-grade and Regents performance turns out not to be useful for evaluating the 8th-grade standards because the cut scores for passing the two basic Regents exams are themselves so low.

The NAEP results for 2009 in this memo are preliminary and are solely for the use of the addressees. As of today, the U.S,. Department of Education has not made available student-level data from the 2009 NAEP. The results presented here are an approximation based on student-level 2007 data and aggregate 2009 data, explained briefly in the Appendix. We will redo the 2009 estimates if student-level data are made available in time. We do not expect any substantial changes, so this memo should be sufficient for planning purposes. If the 2009 data are not made available early enough for the Regents meeting, I will provide a final version of this memo so that the results can be presented to the Regents.

DRAFT: not for circulation except to addressees

National Assessment of Educational Progress

We compared performance on the New York 8th-grade tests and NAEP to answer two questions:

- Where do the New York cut scores fall on the national distribution of performance? This is a measure of the rigor of the standards.
- How did this relationship change between 2007 and 2009? This indicates the degree to which the New York cut score became more lenient, either by score inflation or a failure of linking.

The procedure we followed provides only an approximate comparison of the 8th-grade tests to NAEP. It had several steps:

1. Calculate the percentile ranks (PRs) of the New York cut scores, that is, the percentage of students who scored below each cut score. For example, in 2007, just over 12 percent of New York students scored below Level 2 on the 8th-grade math test, so the PR of the Level 2 cut is approximately 12.

2. Find the scores in the in the New York NAEP sample that are needed to reach the same PRs in those data. In 2007, the New York NAEP score corresponding the 12th percentile rank was 239.

3. Find the PR in the national NAEP data corresponding to the scores obtained in step 2. For example, in 2007, approximately 12 percent of students nationwide scored below 239.

What should one expect from these comparisons? Given the particulars of New York's NAEP performance, one could expect one of three patterns:

1. In the absence of score inflation, one would expect the percentile ranks on the New York test and the national NAEP to be fairly similar (as in the example above). This is explained in the Appendix.

2. If the rapid increase in scores on the New York State tests (more rapid than on NAEP, particularly in math) represents real and generalizable gains in achievement, one would expect the percentile ranks of the performance standards to drop in the New York data but to stay roughly stable in the national NAEP data. That is, the rigor of the standards would be stable, but more New York students would exceed it because of higher achievement.

3. If the score gains on the New York test represents either score inflation or a decline in stringency of the cut scores for other reasons, such as a failure of linking, one would expect the percentile ranks corresponding to the cut scores to drop in both the New York State test data and in the national NAEP data.

On the 2007 New York math test, only 12 percent of students failed to reach at least Level 2, and all but 41 percent reached or exceeded Level 3 (Table 1). However, this does not indicate that a large proportion of New York students were reaching high levels of performance by national standards. As expected, the percentile ranks in NAEP

were similar: the New York standards in effect in 2007 corresponded roughly to the 12th and 36th percentile ranks nationally.

Table 1. Percentile ranks of New York 8th-grade math cut scores,
NY tests and national NAEP, 2007 & 2009

	2007		2009	
	NY	NAEP	NY	NAEP
Level 2	12	12	4	2
Level 3	41	36	20	19

Because of the rapid increase in scores on the New York test, the percentile ranks of the performance standards on the New York tests necessarily dropped dramatically from 2007 to 2009. The percentage of students failing to reach Level 2 dropped from 12 to 4 percent, and the percentage failing to reach Level 3 dropped from 41 to 20 percent (Table 1).

The corresponding percentile ranks in the national NAEP scores dropped by a roughly similar amount. The level of achievement needed to reach Level 2 had fallen so low that only 2 percent of students nationwide failed to reach it (Table 1). Similarly, the performance required for Level 3 dropped sharply, such that only 19 percent of students nationwide failed to reach it.

This shows that by 2009, the Level 2 and Level 3 cut scores were very low in terms of the national distribution of performance. It also suggests that relatively little of the score gain on the New York 8th-grade math test represented real, generalizable improvements in student performance.

The pattern is similar but less extreme when ELA is compared to NAEP reading. In 2007, the percentile ranks of the New York Level 2 and Level 3 ELA cut scores corresponded quite closely to the percentile ranks in the national NAEP data (Table 2). Level 2 was a very lenient standard, while Level 3 was moderate. By 2009, the Level 2 cut score had dropped so far that only roughly 2 percent of students nationwide would have failed to reach it. The Level 3 standard dropped in rigor as well, albeit less so than in math. In 2009, only roughly 34 percent of students nationally would have failed to reach the New York Level 3 cut score. Thus in ELA as well, the apparent improvement in performance on the New York test appears to arise in substantial part from score inflation, a lowering of the standards, or both.

Table 2. Percentile ranks of New York 8th-grade ELA cut scores,
NY tests and national NAEP, 2007 & 2009

	2007		2009	
	NY	NAEP	NY	NAEP
Level 2	6	6	2	2
Level 3	37	45	30	34

Regents examinations

The students who took the 8th-grade test in 2006 are the only cohort for which we have complete data for the Math A Regents exam. Of all the students who took the 8th-grade math test in 2006, 83 percent passed the Math A Regents with a score of at least 65 on their first attempt, and 37 percent scored at least 85 on their first attempt (Table 3).

Table 3. Percent of students passing Math A or Integrated Algebra on the first attempt

	Regents at least 65	Regents at least 85
Math A: total 2006 cohort	82.5	36.6
Math A: marginally proficient, 2006 cohort	90.6	13.2
Integrated Algebra: marginally proficent, 2008 cohort	73.5	0.8

To evaluate the leniency or severity of a given cut score, one needs to consider students who just barely exceed the cut score. This is the group that illustrates the level of performance required to reach the cut score. For this purpose, we used students whose scores were within 4 points of the minimum score for Level 3 on the 8th-grade mathematics test. (This number of points was used in order to have a large enough sample for a stable estimate). We call students who scored within this range "marginally proficient."

Fully 91 percent of marginally proficient students passed Math A with a score of at least 65, but because the 65 cut is so low, this percentage was only slightly higher than in the cohort as a whole (Table 3). (Because the Level 3 cut score is very low, considering only marginally proficient students also removes a large number of higher-scoring students. This accounts for the lower percentage of marginally proficient students scoring 85 or better on the Math A exam.)

We expected a declining pass rate on Math A for marginally proficient students in the 2007 cohort because of the rapid raise in 8th grade math scores. However, we cannot provide comparable tabulations for that cohort because there was a sharp drop in 2007 in the proportion of 8th-grade students who later took the Math A exam. We assume this reflects a decision to wait for the introduction of the Integrated Algebra exam.

We carried out similar tabulations for the Integrated Algebra examination for the first full cohort that took that exam, the students who took the 8th-grade test in 2008. The Integrated Algebra exam appears somewhat harder than Math A for marginally proficient students, with only 74 percent passing on the first try (Table 3). However, because there is no linking of the two Regents exams, this difference cannot be interpreted. It is possible that the drop in the pass rate for marginally proficient students represents inflation of 8th-grade scores or a decline in the rigor of the 8th-grade Level 3 standard rather than a difference in the actual difficulty of the two Regents examinations. That is, if scores became inflated from inappropriate test preparation, or if the linking procedures did not work and the Level 3 standard became more lenient as a result, fewer marginally proficient students would pass Integrated Algebra even if that test was comparable to Math A in difficulty.

Because the Regents ELA exam is most often taken the first time in 11th grade, we could examine the relationship between 8th-grade performance and ELA scores only for the cohort that took the 8th-grade ELA test in 2006. The pattern is similar to Math A: the passing standard for ELA is low enough that the exam is not a useful basis for evaluating the 8th-grade cut scores. Roughly 85 percent of all students with 8th-grade ELA scores in 2006 passed the ELA Regents on the first try, and 94 percent of marginally proficient students passed on the first try (Table 4).

Table 4. Percent of students passing ELA on the first attempt

	Regents at least 65	Regents at least 85
ELA: entire 2006 cohort	84.9	39.1
ELA: marginally proficient, 2006 cohort	94.0	32.1

Appendix

The 2007 NAEP comparisons above were obtained by tabulating student level data from all three databases: the New York State tests, the New York State NAEP sample, and the national NAEP sample.

Percentile ranks on the NAEP and a New York test would be expected to be similar in the first year of administration of the latter because of the New York State NAEP distribution. Step 2 of the process explained above aligns the percentile ranks on the New York State test and the New York NAEP sample. Thus, these percentile ranks are necessarily the same, and the only new information from this step is the NAEP scores that correspond to these PRs in the state NAEP data. Step 3 adds the corresponding PRs in the national NAEP distribution. Because New York's distribution of NAEP scores is very similar to the nation's in both mathematics and reading, the percentile ranks should be very similar.

The 2009 NAEP student-level data are not yet available, although aggregate data are available for the state and the nation. The estimates above for 2009 were obtained by taking advantage of the fact that in both reading and math, the NAEP distributions were very similar in New York and the nation, the changes in the two distributions between 2007 and 2009 were very similar, and the standard deviations were stable. Therefore, mapping 2009 state data to the 2007 NAEP data creates offsetting biases in steps 2 and 3. In step 2, a given PR obtained from the NYS data would map to a lower NY NAEP score in 2007 than in 2009 because state NAEP scores increased from 2007 to 2009. Therefore, using 2007 NAEP scores creates a negative bias. However, for the same reason, in step 3, the NAEP score obtained in step 2 would map to a higher PR in the 2007 national data than in the 2009 national data, creating a positive bias. The net effect of these offsetting biases should be a reasonable estimate of the results that would have been obtained with 2009 student-level NAEP data. This will be confirmed if the 2009 data are made available in time.