Report on Florida's School Grade Scores

Richard Hill
Center for Assessment

August 3, 2010

Background

Schools in Florida get grades based on student performance over 8 components covering several content areas. Four of the components are based on status (reading, mathematics, writing and science) and four are based on the gains students make from one year to the next on Florida's Developmental Scaled Scores (DSS). Gains are measured for reading and math; for both content areas, one gain score is computed for the gains made by all students, and the other is based on the gains made by students in the bottom quarter of the distribution. Schools receive points for each component, which are then summed to total score. The total score gets translated into a letter grade (A-F), although failure to meet certain gain criteria can lower a school's grade.

After 2010 preliminary school scores were sent to districts for their review, many districts raised concern about the drop in grades from last year. For example, over 300 fewer elementary schools earned A's this year compared to last year. As a result, the Florida Department of Education (FDE) hired HumRRO and the Center for Assessment to take a look at the results and attempt to determine whether the declines could be attributed to a processing error or an anomaly in the data or system. This report provides the findings from the Center for Assessment.

Preliminary Information

FDE, after consulting with local district staff, generated a list of several research questions, each of which might have been an explanation for the decline in test scores. After HumRRO and the Center had two weeks to complete their initial review, FDE hosted a meeting attended by several district assessment directors to refine the questions. This process greatly facilitated the Center's review, and provided a local context that was invaluable.

One of the valuable suggestions made by the district assessment directors was to look at changes from the previous year, drilling down in more detail where the changes were the greatest, and then looking at whether those changes were larger than changes found in previous years. Table 1 provides an overall look at how the mean school total scores have varied across the years for different types of schools.

Table 1

Mean School Total Score, by Type of School, for Past Four Years

Type of School	Year				Change between 2009 and 2010
	2007	2008	2009	2010	
Elementary	549	542	562	541	-21
Middle	520	538	545	541	-4
High	483	506	491	496	+5
Combination	514	531	539	528	-11

Clearly, the changes for 2010 were greatest for the elementary schools. Combination schools are those that have grades across the various levels, so their drop is almost certainly consistent with the average decline across the other three types of schools. Note also, however, that the overall scores—especially those for elementary schools—have changed little since 2008. This caused one of the district assessment coordinators to ask which year's data should be questioned—2009 or 2010. That is, if elementary school scores hadn't gone up in 2009, would we have seen a drop this year?

The first step is look at changes in all the components for each of the three major types of schools. Tables 2a, 2b and 2c provide four years' worth of averages for the eight components that comprise the school total scores. The results for all three grades ranges of schools show that results have been more positive for status scores than gain scores, and this is true whether the comparison is back to 2009 or 2008. Table 2a shows that of the 21 points that elementary school grade scores dropped from last year, 4 come from the status scores, and 17 from gain. Therefore, it is clear that declines in gain scores had a much greater impact on changes from last year than did changes in status scores. Interestingly, if we look at changes from 2008, we see that the status scores have gone up 12 points, while the gain scores have declined 11 points. So while some of the explanation for this year's lower scores might be a function of higher scores last year, it is clear that gain scores have been declining, whether compared to 2008 or 2009.

Table 2a

Scores for Each Component for Elementary Schools for Past Four Years

Category of Component	Component	Year				Change between 2009 and 2010	Change between 2008 and 2010
		2007	2008	2009	2010		
Status	Reading	75	75	78	76	-2	+1
	Mathematics	72	74	76	76	0	+2
	Writing	83	80	88	84	-4	+4
	Science	45	47	50	52	-2	+5
	Total					-4	+12
Gain	All Reading	72	66	70	65	-5	-1
	Low 25% Reading	67	63	66	59	-7	-4
	All Math	66	67	67	64	-3	-3
	Low 25%	69	68	67	65	-2	-3

	Math						
	Total					-17	-11

Table 2b

Scores for Each Component for Middle Schools for Past Four Years

Category of Component	Component	Year				Change between 2009 and 2010	Change between 2008 and 2010
		2007	2008	2009	2010		
Status	Reading	63	66	67	68	+1	+2
	Mathematics	63	65	65	67	+2	+2
	Writing	91	92	94	90	-4	-2
	Science	43	43	45	47	+2	+4
	Total					+1	+6
Gain	All Reading	60	64	66	64	-2	0
	Low 25% Reading	69	72	70	71	+1	-1
	All Math	64	66	71	65	-6	-1
	Low 25% Math	67	69	67	69	+2	0
	Total					-5	-2

Table 2c

Scores for Each Component for High Schools for Past Four Years

Category of Component	Component	Year				Change between 2009 and 2010	Change between 2008 and 2010
		2007	2008	2009	2010		
Status	Reading	43	48	47	48	+1	0
	Mathematics	70	74	75	76	+1	+2
	Writing	85	84	82	87	+5	+3
	Science	41	41	40	42	+2	+1
	Total					+9	+6
Gain	All Reading	51	56	51	52	+1	-4
	Low 25% Reading	73	77	75	75	0	-2
	All Math	49	50	48	45	-3	-5
	Low 25% Math	67	71	67	66	-1	-5
	Total					-3	-16

So to this point, it appears as though declines in gain scores, in elementary schools in particular, is the area on which to focus. In the elementary schools, declines in reading gains are

greater than those of math, but that is not true for the middle and high schools, so it raises the question of whether there is something unique in the data for elementary schools that would allow that to happen. Note also that drops are greater for the low quarter of students in reading in elementary schools (indeed, that is the cell with the greatest drop from 2009 to 2010), but that is not the case for the other schools.

Consequently, the next step was to look at gains by grade in elementary schools. One point to note here is that there is not a direct relationship between DSS gains and the gain scores that schools earn. Schools get credit for students who make the required DSS gain, but also for students at Levels 3, 4 and 5 who retain their level from one year to the next. This leads to a source of confusion for people looking at school gains. Students in the lower levels make more likely to post the required amount of gain from one year to the next than students at the higher levels, but students at the higher levels are more likely to earn credit for their school. So when people say that students at the lower levels (in particular, those in the bottom quarter) make smaller gains, that is not a completely accurate statement; students in the lower levels make greater gains, but are less likely to earn credit for their school.

Table 3 provides the *changes* in DSS gains for all students for the past two years, by grade; that is, the entry in each cell is the difference between the amount of gain one year and the comparable gain the previous year. So, for example, students had an average gain of 189 DSS points going from grade 3 in 2007 to grade 4 in 2008, and an average gain of 200 DSS points going from grade 3 in 2008 to grade 4 in 2009. Thus, the students finishing grade 4 in 2009 had 11 points more of gain than did their counterparts a year earlier. To help avoid confusion, we will use the word "gain" to refer to the changes in DSS points from one year to the next, and "change" to mean the difference between two years' worth of data.

One item to note is that the greatest changes were in reading were declines in grades 5 and 6 this year, but that these losses followed gains of similar magnitude the previous year. Another is that the variation in gains is greater for reading than it is for math.

Table 3

Changes in DSS Gains, by Grade

Grade	Reading		Math	
	2008 to 2009	2009 to 2010	2008 to 2009	2009 to 2010
4	+11	-6	-1	-17
5	+19	-36	-14	-20
6	+36	-22	+1	+4
7	-9	+10	-25	-10
8	-2	-10	-19	+12
9	-15	-8	-7	0
10	-26	+14	-16	+3
Mean	2	-8	-12	-4
Standard Deviation	21	17	10	12

In contrast, Table 4 provides the changes in gain component of school grades. As noted earlier, schools get scores for each component in the school grading system. The mean DSS gain of 189 points from grade 3 in 2007 to grade 4 in 2008 led to an average component score of 70; the next year, when the mean DSS gain grew to 200 points, schools received an average score of 73, an increase of 3 points in that statistic. Although there is an obvious correlation between the two statistics, an extreme gain in one does not necessarily result in the same extreme gain in the other. For example, DSS gains at grade 6 increased by 36 points in reading between 2008 and 2009, matched by a 7 point increase in school gain scores. But when there was a 36 point drop in grade 5 DSS reading gains between 2009 and 2010, that led to a 5 point drop in school gain scores, not a 7 point drop. In contrast to the changes in DSS gains, the changes in the school gain scores appear to be no less stable for reading than they do for math.

Table 4

Changes in School Gain Scores, by Grade

Grade	Reading		Math	
	2008 to 2009	2009 to 2010	2008 to 2009	2009 to 2010
4	+3	-3	+2	-3
5	+4	-5	-2	-4
6	+7	-3	+1	+2
7	-1	+1	-4	-2
8	-1	-2	-4	+3
9	-4	0	0	-1
10	-5	+3	-4	+2
Mean	0	-1	-2	0
Standard Deviation	4	3	3	3

To explore the issue of variation in school gain scores by component, we computed the variance of changes across years, by grade, for each of the six reading and math components (status, gain for all students, gain for the bottom 25 percent), averaged those variances across all grades, and the computed the square root, in order to obtain an estimate of the standard deviation of the change for each component. The results are reported in Table 5.

Table 5

The Standard Deviation of Changes in School Grade Scores, Reported by Component

Content Area	Component		
	Status	Gain for All	Gain for Bottom 25%
Reading	2.2	4.0	5.5
Mathematics	1.7	2.6	3.7

Table 5 provides a coherent picture of the nature of change in school grade score components; reading varies more than mathematics from year to year, the gain component varies more than status, and gain for the bottom 25 percent varies the most.

So, in summary, we know the following about recent history of changes in Florida's school grade scores:

1. Status scores have improved while gain scores have declined when compared to 2008 for all three categories of schools.
2. Declines in school grade total scores were greatest this year in elementary schools. However, last year, the biggest drop was in high schools, so there is no consistent pattern there.
3. When comparing to 2009, the declines in gain scores have been greatest in elementary schools, but when comparing to 2008, the declines in gain scores have been greatest in high schools.
4. Declines in gain scores have been approximately equal for both reading and mathematics.
5. When looking at change from year to year, the variation has been greater in reading than in math, greater for gain scores than for status scores, and greater for the gains of the bottom 25 percent than gains for all students.
6. The single area of greatest decline is gain scores in reading for elementary schools, particularly grade 5.

Normal Random Variation

Before we start trying to identify the reasons for the changes in scores that have taken place between 2009 and 2010, it is useful to determine how much of the changes could be due simply to normal random variation. For example, the equating that is done from year to year is produced from a sample of questions to be anchor items. Choosing another set of anchor items would alter the results somewhat and cause the results to vary from what they were.

The amount of normal random variation, and the sources that significantly contribute to it, vary across the reporting level. For example, at the state level, student sampling is a trivial factor, since the state tests hundreds of thousands of students at each grade each year. But at the school level, the set of students tested in any particular year can play a significant role in the results—there truly is a "good class, bad class" issue that needs to be taken into account when evaluating school results. On the other hand, given a set of students in a school, the occasion of testing ("good day, bad day") rarely plays a role, since those events tend to average out across all the students in a school—but occasion is a significant source of error in the interpretation of student results. On the other hand, the choice of items for equating, which is a significant factor at the state level where other sources of variation are very small, is a trivial factor at the student level, where occasion can play a major role. So it is critical to think about the unit of analysis when deciding which factors of variation need to be controlled.

So how much variation in state results should be expected? The data in Table 7 shed some light on the issue. Table 7 provides the changes from the previous year in the state mean DSS.

Table 7

Changes in Statewide Mean DSS from Previous Year

Content Area	Grade	Year					Mean	SD
		2006	2007	2008	2009	2010		
Reading	4	-28	11	19	28	-4	5	22
	5	8	28	-23	33	-8	8	24
Math	4	25	6	22	23	2	16	11
	5	1	13	9	8	3	7	5

Remember that we chose the data in Table 7 because these were the grades in the schools that had the greatest decline in gain scores this year. With that in mind, Table 7 provides some interesting information:

1. Reading scores and math scores in these two grades have been going up at about the same overall rate, but a little faster in math than in reading (which is not atypical).
2. The standard deviation of gains is substantially larger for reading than it is for math.
3. For these two grades, the gains from 2008 to 2009 were higher for reading than at any other time in the five-year look provided by this table. So the area that we chose to focus on because it had the greatest decline in gain scores in 2010 had its greatest gains in five years the previous year.

Table 8, which is an extension of the data already shown in Table 3, develops this point further. The data in Table 8 are not the same as those in Table 7. Table 7 provides the changes in the DSS status score from year to year; so, for example, the value of -28 in the first cell in Table 7 reflects the fact that the state's mean DSS for grade 4 in 2005 was 1575; in 2006, the state's mean DSS had dropped to 1547. Table 8, on the other hand, provides gain scores; the entry of 189 in the first cell reflects a change of 189 DSS points for students who had matched data in grade 3 in 2007 and grade 4 in 2008. Table 8 provides the mean change in DSS, aggregating up from student matched scores from one year to the next. In contrast to Table 7, it shows more grades but fewer years. But the results are almost identical—standard deviations of around 20 for reading, and about half that for math.

Table 8

Mean DSS Gains and Changes, by Year and by Grade

Content Area	Grade	Year			Change from 2008 to 2009	Change from 2009 to 2010
		2008	2009	2010		
Reading	4	189	200	194	+11	-6
	5	63	81	46	+18	-35
	6	67	103	81	+36	-22
	7	113	104	115	-9	+11
	8	95	92	82	-3	-10
	9	79	64	56	-23	-8
	10	41	15	29	-26	+14
	Mean				2	-8
	Standard Deviation				21	17
Mathematics	4	112	111	94	-1	-17
	5	132	118	99	-14	-19

	6	35	36	40	+1	+4
	7	155	130	120	-25	-10
	8	94	75	87	-19	+12
	9	60	53	53	-7	0
	10	56	41	44	-16	+3
	Mean				-11	-4
	Standard Deviation				10	12

Whether we look at a grade across years, or look across grades within a year, we get the same result: the standard deviation of reading gain scores is around 20 points and the standard deviation of math gain scores is around half that. So the variations we are seeing in Florida's scores this year are not outside the normal range. And when we look at the average change across grades in a year, we see that the trends are quite small. When we focus on the areas that have changed the most this year (such as grade 5 reading), we are apt to over-interpret the data and come to conclusions that are not warranted when looking at the bigger picture.

Responses to External Review and District Test Coordinator Questions

It is clear from the discussion in the previous section that results vary from year to year. The questions that have been posed all focus on a central issue: Do we believe the variation this year is part of the natural process, or due to some more subtle statistical anomaly? The answer is not going to be the same for reading as for mathematics. For reasons to be explained later, there is more variation in the equating of reading scores across years than for mathematics (which likely is why we saw, in the section above, that there is more variation in reading results from year to year than there is for mathematics). Thus, answers to the questions that have been raised are more unambiguous for mathematics than for reading. Where necessary, we will discuss the results for both content areas, but when looking for subtle statistical explanations, we will look primarily at mathematics results (until the section in which we discuss equating).

Several possible reasons have been suggested for the drop in scores, particularly elementary schools, this year:

1. Test construction specifications, and following those specifications
2. Composition of students
3. Field testing in reading
4. Equating specifications, and following those specifications

The remainder of this report will include our findings on each of these issues.

Test content specifications

FDE will produce its own report on this issue. We have reviewed the test content specifications and found them sufficiently explicit and unchanged across the years. We also believe that the tests have been consistent drawn from these specifications, and therefore consider that test construction has not been a factor in test score change. However, the district assessment coordinators raised more questions than a person without inside knowledge could expect to answer, and therefore FDE will address these points in a report of its own.

Composition of Students

Table 3 showed that DSS gains in mathematics declined, on average across the grades, both from 2008 to 2009, then again from 2009 to 2010, although the average declines were small, and smaller for the change from 2009 to 2010 than the previous pair of years. Two related questions have been raised, and they deal with the gains of the bottom 25 percent of students: how have the gains of that subgroup changed, and to the extent there are differences from the general population, could those differences have come from a changing composition of that group? Because the answer to these questions will involve subtle differences that require very solid equating from year to year, we will answer it using the mathematics data only.

First, remember that there is a difference between making DSS gain and receiving credit for making DSS gain; students can receive credit for making gain if they are in Levels 1 and 2 only by making the required amount of DSS gain or by increasing their level. Students in Levels 3-5 can receive credit for making gain if they make the required amount of DSS gain or maintain the level they were in the previous year. Thus, while students at the lower levels make, on average, more DSS gain than students at higher levels, they are less likely to earn credit for their school. Therefore, the answers to questions about gain, especially as they relate to school grade scores, are often quite complex. To start, however, we will look simply at changes in mean DSS scores.

Table 9 provides the changes in DSS gains for each level of students in the bottom 25 percent, along with the average across all students, by grade for the past two years. It is hard to discern much of a pattern in the data, except that gains for the Level 1 students dropped dramatically from 2008 to 2009, but stayed stable from 2009 to 2010. Given the level of concern over the results this year that did not exist last year (at least to this degree), that is a somewhat surprising result. But it is not true that there is a general pattern to find great gains after a year of great loss, or vice versa. While that did happen at certain grades, there are other grades that provide a counter-example. And while the gains of the bottom 25 percent dropped more than all students between 2008 and 2009, the opposite was true between 2009 and this year.

Table 9

Changes in DSS Gains across Consecutive Years,
For Students in Bottom 25 Percent, by Level, and All Students

Grade	Pair of Years	Level			All Bottom 25 Percent	All Students
		1	2	3		
4	2009 - 2008	-43	8	10	-16	-1
	2010 - 2009	1	-7	-15	-9	-17
5	2009 - 2008	-21	-4	-9	-14	-14
	2010 - 2009	-15	-17	-16	-22	-20
6	2009 - 2008	+15	+5	-3	+1	+1
	2010 - 2009	0	+3	+1	+10	+4
7	2009 - 2008	-66	-13	-8	-59	-25
	2010 - 2009	-16	-2	-6	-14	-10
8	2009 - 2008	-41	-9	-9	-30	-19
	2010 - 2009	+40	+14	+3	+31	+12
9	2009 - 2008	-33	-5	-1	-26	-7
	2010 - 2009	-16	-7	-0	-10	0

10	2009 - 2008	-53	-17	-15	-36	-16
	2010 - 2009	+3	+7	+10	+4	+3
Average across All Grades	2009 - 2008	-35	-5	-5	-26	-12
	2010 - 2009	0	-1	-3	-1	-4

Another look at the gains of the bottom 25 percent is provided in Table 10. This table provides, for the past three years, the composition of the bottom 25 percent, as well as the likelihood that a student at each level will make the required DSS gain and the likelihood that they will earn credit for their school. The data show that the composition of the bottom 25 percent has indeed changed somewhat over the past three years; there are more students in Level 3 in that group than was the case in earlier years. And while students are Level 3 are less likely to make the required amount of gain than students at either Level 1 or 2, they are more likely than students in Level 2 to earn credit for their school. But of more concern is the fact that students at all three levels are less likely to make the required gain now than they were two years ago.

One can calculate what the weighted average for scores would be if the composition of the bottom 25 percent had not changed from 2009 to 2010. In 2010, 63 percent of the bottom 25 percent students received credit for making gain (the weighted average of .72*.33 + .58*.40 + .60*.27). If the proportions of students at each level had remained what they were in 2009, the weighted average would have been 64 percent (.72*.37 + .58*.40 + .60*.23). The actual figure for 2009 was 67 percent. Thus, 1 percentage point of the 4 percentage point change could be attributed to the change in the composition of the levels of students in the bottom 25 percent. Thus, while it was a factor, it played only a small role in the decline of scores across the two years.

Table 10

Gains Made by Students in the Bottom 25 Percent

Level	2009-2010			2008-2009			2007-2008		
	Gain	Credit	% of Pop'n	Gain	Credit	% of Pop'n	Gain	Credit	% of Pop'n
1	.70	.72	33	.71	.73	37	.73	.76	41
2	.57	.58	40	.61	.62	40	.61	.62	39
3	.43	.60	27	.49	.65	23	.49	.64	20

Field Testing in Reading

A new type of reading item was field tested this year. The district assessment coordinators reported that this created administration problems in some locations, and questioned whether that might have been a factor in the declines of reading scaled scores.

It turned out that the anchor items were taken by a representative sample of students at certain grades. In the past, anchor items have been taken by students only in selected "early return" schools, but this year, the anchor items were spiraled with all the other forms of the test in grades 3, 4 5 and 9, and thereby taken by a representative sample of students over all schools. Student who took the anchor items did not receive the new reading field test items, so one way of determining whether the administration of the field test items led to lower test scores was to compare the performance of

the students who took the anchor items (but not the field test items) scored any higher or lower than the students who took the field test items (but not the anchor items).

Table 11 provides the effect sizes for the performance of students taking the anchor items versus those taking the field test items. Note that the administration of the field test items affected reading only, so differences in mathematics performance gives a sense of the amount of random variation one should expect in these results.

Table 11

Effect Sizes for Students Taking Anchor Items vs. Students Taking Reading Field Test Items

Grade	Reading	Math
3	+.00	+.01
4	+.00	+.04
5	+.04	-.01
9	+.07	+.07

With the exception of Grade 5, there is no evidence that the administration of the field test items affected student performance—and if there was an effect in Grade 5, it was quite small.

Equating Specifications

The following is a brief conceptual overview of how equating works in Florida. The DSS scale was developed a decade ago. The tests given each year consist of core items (questions which are given uniformly to every student within a grade and count toward the student's DSS score) and either field test items or anchor items. The anchor items are those which have been given in previous year, have been linked to the DSS scale, and are given to a sample of approximately 4,000 students. There is no assumption in the equating design that sample is representative of state, but data shows that the samples drawn are invariably close to the state average. The model compares how those students do on the core items versus the anchor items. This comparison tells us the relative difficulty of this year's core items to the anchor items, and knowing the location of the anchor items on the DSS scale, thereby provides the information necessary to place the core items on that same scale.

There are several issues of importance when looking at Florida's equating practices:

1. First, and perhaps most importantly, Florida uses techniques and specifications for its equating that reflect the highest standards of current practice. The techniques and specification have been reviewed by outside agencies of high repute, and reflect much refinement over the years. Just as importantly, there is ample documentation that the specifications have been followed to the letter.
2. Florida has three independent groups replicate all its equating; one within the Department, and two outside agencies. All discrepancies occurring among any of them is reviewed by all, so there is great assurance that operational errors do not occur.
3. As we have noted in earlier sections of this report, there is random variation that occurs with any equating, mostly based upon the particular choice of equating items used in any particular year. That variation is greater for reading than for mathematics, because

although the same approximate number of items are used for both content areas, the reading items are passage-related and therefore not independent of each other. The variation in equating goes down when the equating items are each independent estimates of change across years. It also is natural that there is more variation in gain scores than status scores, since gain scores incorporate two years of variation while status scores only have one.

4. The various equating items available to be used in any particular year have been linked to a particular year in the history of the program. Sometimes, the equating items are drawn from the previous year; when that happens, that provides a *direct link* to the previous year. When they are drawn from an earlier year, however, they provide a direct link to that earlier year, and are related to the previous year by a process of *chaining*, where one uses a series of links to relate the two years. There is less random variation for a direct link than for a chain. Thus, if one is measuring gain, the variation due to equating is least when one provides a direct link to the previous year. In contrast, if one is measuring status (and changes in status since the establishment of the baseline), the variation due to equating is least when one provides direct links as close as possible to the base year and a minimum of chaining is done.
5. When equating, one is most certain of the accuracy of the results when item position is held constant. While it often will not matter if items are moved a bit from what their position in the test was when the parameters for the items were established, items sometimes unexpectedly will be affected significantly when moved. This tends to happen more often for reading items than math items. Unfortunately, because the reading items are passage related, their item position will change more frequently than is true for math items. So we often have to move the items we would wish to move the least. Florida attends to this issue, and maintains item position as closely as possible within the constraints of test development and administration.

If one assumes that real progress in Florida has been fairly constant from year to year, then the standard deviation of the gain scores defines "normal random variation." It is greater in reading than in math because, while both content areas use about the same number of items for equating (around 30), the math equating items are all independent of each other, while the reading equating items are passage-related (four passages are typically used) and therefore, inter-dependent—meaning that each additional reading equating item does not provide as much additional equating information as would be realized if all the items were independent. Thus, while the average gain in grade 4 reading has been 5 points, the actual reported gain for any particularly year, relative to the previous year, has often been very different from that. Note the contrast to grade 5 math, where the reported gain has been much more consistent. The greater reliability of the math equating process (because of the ability to use independent items) is almost certainly the reason for that smaller standard deviation.

So under the current equating model, we can expect greater fluctuation in reading scores from year to year. Looking specifically at 2009, we saw unprecedented reported gains in scores in both grade 4 and grade 5 that year. What to expect in 2010 was a function of linking and chaining, or more specifically, knowing what year 2010 scores were directly linked to.

As previously stated, there were four passages used to link 2010 to the DSS scale; three of those passages came from 2008, and the other was from 2009. So three-fourths of the information about where to base the 2010 grade 5 reading scores came from 2008, not 2009. If all the equating passages had come from the 2009 test, then whatever random inflation may have been carried in those scores would have carried over to the 2010 scale, and we might well have seen higher DSS

scores in 2010 than have been computed. But because three-fourths of the judgment about how students scored this year is derived by comparing this year's students to those of 2008, not those of 2009, we have a better look at gain from 2008 than we do from 2009.

Summary

Florida is justifiably proud of its testing program. Lessons have been learned from previous years' problems, and as a result, the current program is well specified. Importantly, those specifications are met throughout the entire process of test development, equating and reporting. Florida's equating reflects the current best practice and advice of leaders in the measurement community, leading to valid tests and test scores from year to year.

Every year, because of normal random fluctuation in the system, some grades show gains in scores while others show declines. This year, the declines in school grade scores happened most in adjacent grades, which meant that the declines this year affected elementary schools far more than others. While there may well be some adjustments that could be made in the equating process to reduce normal random variation even further, the simple fact is that very small changes, particularly in gain scores, lead to noticeable changes in school grade scores, especially when aggregated across large numbers of students and schools. There may be an opportunity to improve on these practices as the next generation of assessments and accountability is phased in in Florida, but for now, what Florida has been doing is based on the current state-of-the-art.

Individual student scores have a large amount of random variation in them. This is true not only for FCAT, but any standardized achievement test will have limitations on interpretation based on the limited sample of student achievement on which these tests are based. Every responsible testing program will tell parents and teachers to use the results of an individual student with caution, because the measurement error associated with the results is large enough to warrant such caution. The normal variation due to equating that has been discussed in this report is a very small number compared to the standard error of measurement associated with individual student reports. So, despite any concerns that might be raised about the validity of score interpretations when the data are aggregated over large numbers of students, parents and teachers should have no concern (beyond the usual) about using and interpreting these FCAT results.

As data get aggregated to higher levels, the concerns about individual student measurement error decrease dramatically, because these errors tend to average out across students. However, in contrast to student level results, small changes in school and district results become more noticeable, and the types of issues such as normal random variation in equating take on an importance that they do not have at the student level.

It does not appear that there was any operational error in the program this year, nor are there simple explanations about changes in the student population that would explain much of the decline seen in elementary schools. It does appear that the usual amount of random variation that has been true every year was replicated this year, and it unfortunately occurred at the grades where the largest number of schools would be affected.