RESEARCH BRIEF
Research Services

Vol. 1002
August 2010

Dr. Terry Froman, Research Services

From Normal Variation to Unacceptable Uncertainty

A Brief Review of Anomalies in the 2010 FCAT Results

The release of School Grades by the State of Florida in the late summer of 2010 created quite a stir. Over 500 elementary schools all over the state dropped at least one grade level. Among those schools that experienced drops, the Reading Gains and Mathematics Gains components seemed especially low. Some of the historically best performing schools saw unprecedented shortfalls in gain percents. Ad hoc analyses of the summary results by districts throughout the state uncovered radical shifts in points earned by schools. The results seemed that they just could not be true. Coordinated calls for review by Superintendents and research staff led the State Department of Education to call for two independent audits for further evaluation.

The findings of the audits supported the initial results and concluded that the fluctuations in gains were within historical ranges. No violations of protocol were uncovered and the School Grades were deemed correct. Despite the authority of the reviewers and assurances from the State, many consumers of the test results at all levels had reservations. The practical results simply did not seem believable. The potential consequences to all concerned could not be taken lightly. Future challenges to the results seemed inevitable and lingering suspicions would likely plague the accountability system for some time to come.

The purpose of this paper is to review some of the FCAT score results that prompted concern, clarify in nontechnical terms the test construction and equating justifications, and propose a perspective for confronting the compromises to accountability in the future.

Understanding Gains

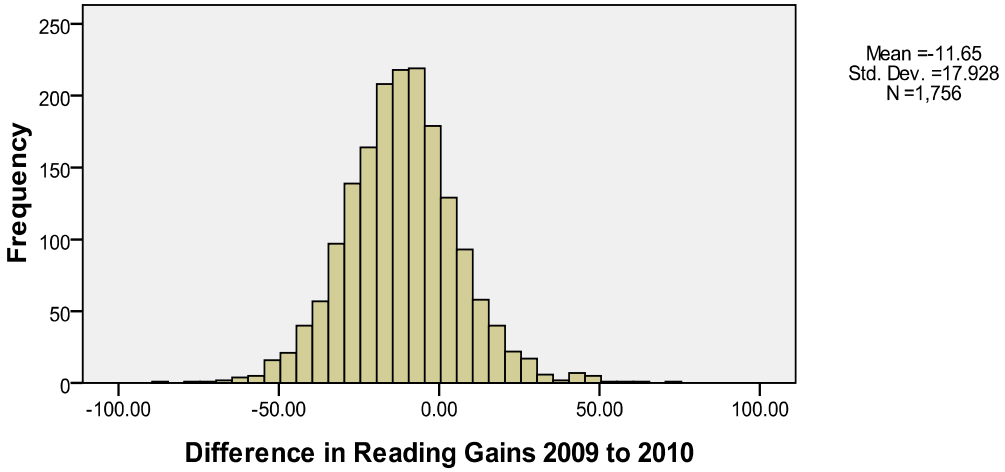
The Reading and Mathematics tests are given in each grade level from grade 3 to grade 10. Schools earn one point for each percent of students making gains. Gains are defined by improving in achievement levels, maintaining at high achievement levels, or demonstrating normal growth in lower achievement levels. Gains among students in the lower 25% are counted again as a separate group. These four gain components – gains among all students in Reading, gains among all students in Math, gains in the low 25% in Reading, and gains in the low 25% in Math – constitute half of the potential points a school can earn toward its grade.

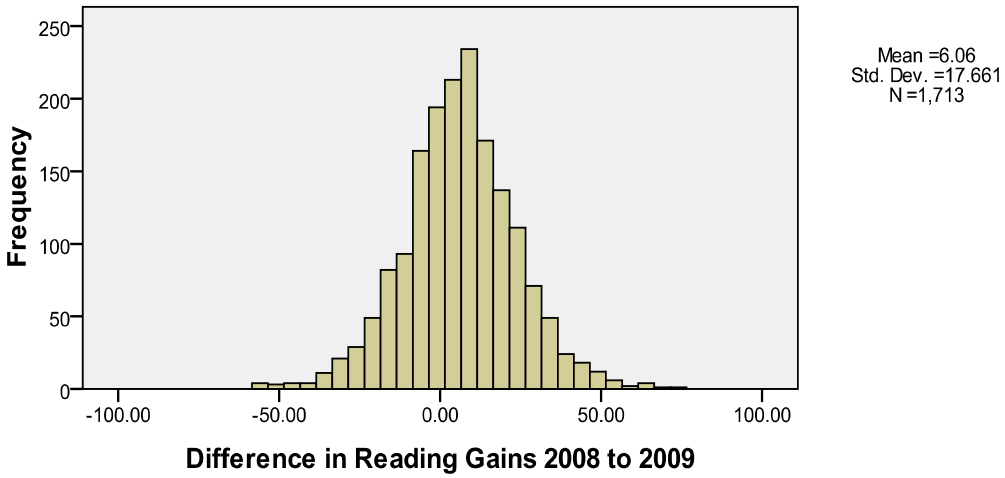
Gains in this kind of grading system are not simply the improvement in developmental scale scores from one year to the next. Simple gain scores, although useful, are notoriously unreliable. Because they are the difference between test scores from two different tests over two separate occasions, they tend to combine the unreliability of the individual tests. The gains in the school grading system, calculated in a more complex fashion but still involving underlying simple score changes, share some of the unreliability problems. This means that, even under well-controlled testing conditions, it should not be surprising to see occasional large changes in gain percentages among schools from one year to the next.

Observed Differences in Gains

While acknowledging the inherent volatility of gain percentages, the observed changes in gains for 2010 were of a surprising nature. Each district readily found extreme cases, but unusual trends confined to one district would not necessarily set off alarms. They might be explained by local policy changes, shifts in instructional emphasis, changes in staff, and other short-range circumstances. However, it quickly became clear that the changes in gains this year were widespread throughout the state.

To get some idea of this, consider the graphs below of all elementary schools in the state presenting the differences in Reading gains between the last two years compared to the two previous years. In these graphs, the gains in regular Reading and gains in the lower 25% Reading are combined. It can be seen that the schools suffered an average loss of over 11 points in combined Reading gains in 2010, after having enjoyed an average increase in the same kind of gains of 6 points the previous year. This means that the average total points per school were different by more than 17 points over that time span – more than enough to bring many schools down one grade level.

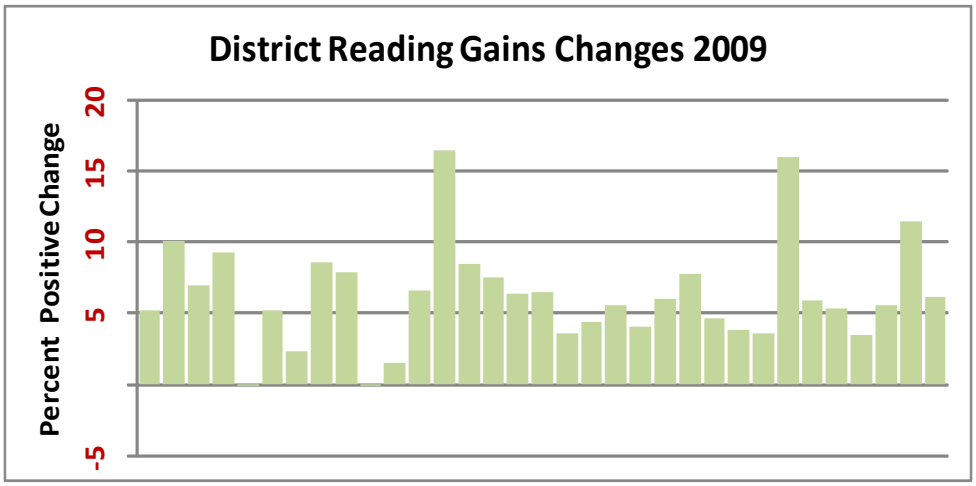


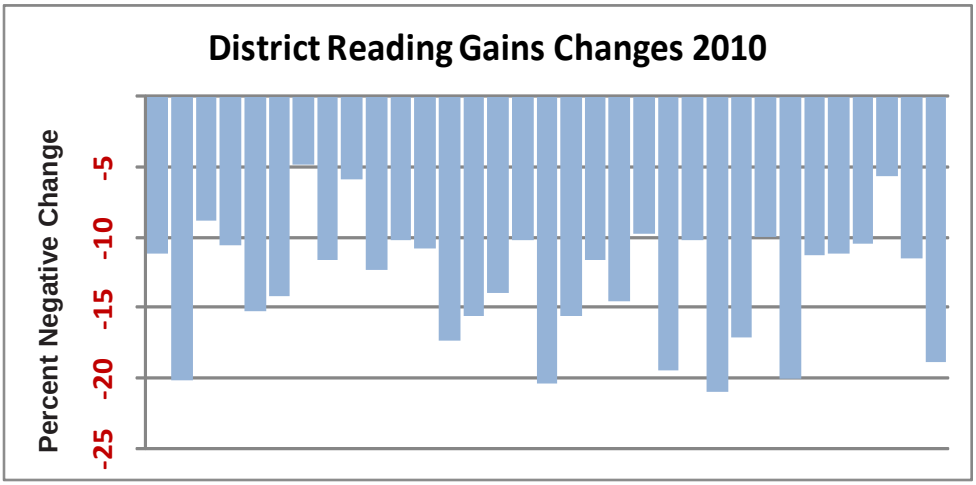


Systematic Change

Even this kind of pervasive difference in gains is not, in itself, an indictment of the test. It may have been that some schools increased and some decreased, varying from district to district, with an acceptable degree of overall state-level change. What would be more persuasive would be consistency in change across districts. Consider the graphs below depicting the elementary school differences in combined Reading Gains for 2009 and 2010 aggregated by district.

In these graphs, the districts are arranged in the same order so that you can compare the changes district by district. Here we see a remarkable degree of consistency across districts. In 2009 almost all districts experienced increases in Reading gains from 5 to 10 percentage points. In 2010 the trend reversed and all districts had decreases of even larger sizes. The regularity in direction and size of change among all districts would strongly suggest that some kind of state-level effect was in operation. The candidates for this kind of effect are limited. It seems unlikely that economic strain or learning growth limits, for instance, would have this kind of systematic consequence. The most likely explanation lies in attributes of the test, itself, affecting all districts in a similar manner.





Test Score Distributions

What are the kinds of changes in scores on individual tests that could contribute to the discrepancies in gains? On the following page are tables showing selected changes in the percentage of students scoring within the five FCAT achievement levels over time for all districts in the State. The first of the three colored tables shows the differences between 2009 and 2010 in the percent scoring at each level for the 4th grade Mathematics test. The differences are color-coded – increases in percent colored green and decreases in percent colored red. So, for instance, in the first line of the first table Alachua County's 4th grade Mathematics percentages changed between 2009 and 2010 with 4% (rounded) more students in Level I, 1% more students in Level II, 2% fewer students in Level II, and 1% fewer students In Levels IV and V, respectively. The pattern of red and green cells throughout this first table is what one might expect to see – slight changes arbitrarily distributed across all districts.

In the middle table, showing changes between 2009 and 2010 in percents within levels for the 4th grade **Reading** test, the situation is very different. Here, one can see surprisingly consistent color-coding down columns, indicating that almost all districts experienced changes in scores in the same fashion. The percentages in Levels I and V generally increased and percentages in the middle levels decreased by compensating amounts. Once again, this consistent pattern suggests that we are looking at some kind of problem assigning scaled scores at the State level that affects all districts uniformly.

The third colored table presents changes in percents within levels for the 4th grade Reading test for **the previous year**, between 2008 and 2009. The consistent red stripe of strong negative entries in Level I down all districts is especially eye-catching. Lower percentages in Level I would be welcome changes within any single district. Perhaps because these changes might have been greeted with satisfaction at the local level, they went unquestioned. However, they were probably the precursors to the patterns of questionable accuracy that were observed in 2010. Because many of these unusual changes in percents involve achievement Level I, it is easy to see how this might have an effect on Reading gains. Students moving from Level I to Level II between years would constitute a large proportion of the recorded gains at any grade level.

This raises the larger issue of when and to what degree these kinds of dramatic score range changes have taken place in the past. There is, of course, the infamous situation a few years ago concerning 3rd grade Reading scores. After exhibiting slight 1 to 3 percent increases historically, the percent of 3rd graders scoring proficient in Reading shot up 8 percentage points in 2006 and then dived 6 percentage points down again in 2007. The company contracted to administer the FCAT was quoted as standing by the validity and reliability of the 2007 scores. It was finally determined that the problem was apparently in the anchor items for 2006 and, after imposed rescoring, the drops in 2007 dissipated.

Overshadowed by the changes in Reading gains in 3rd grade were the equally remarkable changes in Reading gains in 7th grade during that same period. In 2006, 86% of the middle schools throughout the state had increases in Reading gains (average +7 percentage points). This was followed, in 2007, with 87% of the middle school 7th grades suffering losses in Reading gains (average -7 percentage points). It is reasonable to assume that similar size swings in score distributions happened in other years and in other content areas.

			Grade 4 Mathematics 2009 to 2010					Grade 4 Reading 2009 to 2010					Grade 4 Reading 2008 to 2009				
		Levels	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
01	ALACHUA		4	1	-2	-1	-1	2	0	-3	0	3	-5	2	3	-1	1
02	BAKER		3	0	1	-3	-1	5	0	-1	-7	4	-10	-2	6	5	1
03	BAY		-1	0	0	2	0	5	0	-4	-3	1	-4	2	4	-2	1
04	BRADFORD		1	4	3	-7	-1	5	-5	-2	1	2	-9	11	4	-6	-1
05	BREVARD		0	-1	0	0	0	4	0	-4	-3	2	-3	-1	5	-1	0
06	BROWARD		-1	-2	1	0	0	3	0	-4	-2	2	-5	-1	4	2	1
07	CALHOUN		-1	-5	-1	4	3	5	-3	-5	-4	7	-2	-6	8	3	-3
08	CHARLOTTE		-1	-2	-4	4	2	6	0	-2	-4	1	-3	-1	4	0	0
09	CITRUS		-2	1	5	-4	-2	1	0	-1	0	-1	-1	2	2	-3	1
10	CLAY		0	-2	1	0	0	1	1	-4	-1	3	-1	0	4	-1	-1
11	COLLIER		0	1	-1	-3	2	2	-2	-2	2	2	-3	0	1	-1	1
12	COLUMBIA		1	-1	2	0	-1	7	-1	-4	-3	0	-2	3	0	0	0
13	MIAMI DADE		0	-1	2	-1	0	0	-1	-2	0	3	-3	0	3	2	-1
14	DESOTO		1	-8	3	2	2	4	0	-4	-2	1	-3	-4	3	2	1
15	DIXIE		0	3	-4	0	2	7	2	0	-5	-4	-14	2	5	1	7
16	DUVAL		1	-1	1	0	-2	4	0	-3	-2	2	-4	0	4	0	0
17	ESCAMBIA		1	-1	2	-2	1	4	-2	-3	-2	3	-6	3	4	0	-2
18	FLAGLER		0	0	-1	2	-1	1	1	-2	-2	2	0	-1	2	-1	-1
19	FRANKLIN		2	0	10	-9	-4	9	-3	-4	-4	3	-17	-1	11	2	4
20	GADSDEN		-2	-3	8	0	-2	4	-4	-4	4	0	-9	4	5	0	0
21	GILCHRIST		5	9	4	-17	0	1	-5	0	2	1	1	9	-1	-7	-1
22	GLADES		-2	1	2	-3	3	5	-1	-13	9	1	-3	-5	-2	6	3
23	GULF		-1	-1	8	-4	-2	-5	-1	2	5	-1	0	4	-9	7	-2
24	HAMILTON		2	-8	2	1	4	14	-4	-6	0	-3	-11	4	11	-4	1
25	HARDEE		-3	-3	6	1	0	0	-5	1	0	2	-8	0	5	1	3
26	HENDRY		-1	-1	1	3	-1	1	-1	-9	7	2	-1	-5	8	-3	0
27	HERNANDO		1	-1	0	2	-3	2	1	-1	-4	3	-4	0	3	0	1
28	HIGHLANDS		2	-2	-1	0	2	3	0	-1	-3	1	-4	1	1	1	1
29	HILLSBOROUGH		1	0	0	0	0	2	0	-2	-2	1	-4	0	2	2	1
30	HOLMES		1	-4	5	-7	4	1	3	0	-1	-3	0	-5	2	-1	4
31	INDIAN RIVER		-1	2	-3	0	2	1	1	-3	-1	1	-3	-1	3	1	2
32	JACKSON		-2	-3	1	0	3	6	-3	-2	-4	3	-7	-1	4	4	1
33	JEFFERSON		-4	-1	3	5	-2	20	-11	-8	-2	2	-9	10	7	-4	-5
34	LAFAYETTE		-1	-7	11	-13	10	9	4	-10	-7	4	-7	5	10	-4	-4
35	LAKE		1	0	1	-2	1	4	1	-3	-2	1	-5	0	4	0	1
36	LEE		1	0	-3	1	1	4	0	-1	-3	1	-6	0	2	3	1
37	LEON		0	-2	0	1	0	2	-1	0	-3	2	-3	0	1	0	3
38	LEVY		-3	1	0	1	0	8	0	0	-9	1	-9	1	5	4	0
39	LIBERTY		-3	15	-4	0	-7	8	-5	-6	3	-1	-6	-3	12	-4	2
40	MADISON		7	-1	-2	-2	-2	-1	-3	-4	7	0	-2	4	-1	0	-1
41	MANATEE		-2	-1	2	1	0	5	2	-2	-5	1	-6	0	3	3	-1
42	MARION		1	0	1	-3	1	3	0	-3	-3	2	-3	0	2	2	0
43	MARTIN		2	0	-2	0	0	0	1	-5	0	3	-1	-3	5	0	-1
44	MONROE		3	2	-1	-3	1	0	1	-4	0	3	-4	-3	9	0	-2
45	NASSAU		0	3	1	-1	-3	2	0	-1	-4	3	-4	1	0	2	0
46	OKALOOSA		-1	-1	0	2	2	4	0	0	-5	2	-2	0	-1	2	1
47	OKEECHOBEE		2	-5	-1	2	1	5	2	-9	1	1	-7	-1	6	0	1
48	ORANGE		2	-1	1	0	-1	1	0	-2	-2	2	-4	0	2	2	0
49	OSCEOLA		2	-3	0	1	0	3	-1	-4	0	2	-6	0	3	2	0
50	PALM BEACH		0	-2	1	0	1	2	0	-2	-1	2	-5	0	4	1	0
51	PASCO		1	1	-1	0	-1	5	0	-3	-2	0	-5	0	3	0	1
52	PINELLAS		-1	-1	0	1	0	3	1	-3	-1	0	-4	0	4	-1	1
53	POLK		-1	-2	4	1	-1	4	0	-4	-2	1	-5	1	4	0	0
54	PUTNAM		3	0	-5	-1	2	2	4	-3	-2	-1	-5	-4	8	1	1
55	ST JOHNS		0	0	0	-2	1	3	-1	-1	-2	2	-4	0	-1	3	1
56	ST LUCIE		2	0	0	-2	-1	3	0	-5	-1	2	-4	-1	5	0	0
57	SANTA ROSA		0	-1	1	-1	0	3	0	-5	-1	4	-2	1	4	-1	-2
58	SARASOTA		0	-1	0	-1	1	1	-1	-1	-2	3	-3	1	1	0	0
59	SEMINOLE		0	0	-1	1	-1	3	2	-3	-2	1	-4	-2	3	2	1
60	SUMTER		0	1	0	-2	2	0	0	5	-7	3	-1	-5	0	4	3
61	SUWANNEE		1	1	4	-5	-1	8	-3	-6	-3	2	-7	7	3	0	0
62	TAYLOR		1	4	4	-6	-2	-1	0	-3	1	3	-1	2	3	-1	-2
63	UNION		3	0	-1	0	-2	1	2	-2	-5	3	-2	-1	3	-1	1
64	VOLUSIA		0	-2	1	0	0	4	0	-4	0	1	-4	0	3	0	0
65	WAKULLA		1	3	-2	-2	1	4	1	-5	-3	2	-3	2	3	-5	4
66	WALTON		-1	-1	2	0	0	4	-1	-1	-7	5	-6	-1	4	2	2
67	WASHINGTON		1	-2	-1	1	2	1	4	-4	-1	1	-2	-5	-2	8	0

Test Scaling and Equating

People often speak of the FCAT as if it were one test. In fact, it is 22 different tests. There is a unique test for each of eight grade levels in both Reading and Mathematics, and a unique test for each of three grade levels in both Writing and Science. These 22 tests change every year. A certain number of items, the so-called "anchor items," are carried over from year to year to allow for equating and scaling, but the majority of items on all tests are original each year. Creating this many new tests every year, and having each conform to content coverage demands and item characteristics, is a great strain on the managers of the FCAT tests. Of course, blatant lapses in construction or scoring (including potential mismatches in student identification numbers) could have catastrophic consequences for the consumers of these tests, and therefore, great care is taken and the procedures are carefully monitored. However, even under the best control conditions it seems inevitable that slight deviations from ideal test characteristics would creep into the process from time to time. On rare occasions, these aberrations might grow big enough and coalesce unfavorably enough to create a worst-case scenario of misevaluation. This may have been what was experienced with the FCAT tests of 2010.

The methodologies used by the State for equating and scaling the FCAT tests are among the most sophisticated and respected in the field of psychometrics. Although quite complex, at a basic level they rely on a very simple idea relating item difficulty to student ability. Briefly stated, it is presumed that students with higher ability will have a greater probability of getting any individual item correct, and items with less difficulty will have a greater probability of being answered correctly by any individual student. Analysis and construction of the tests are based on mathematical methods derived from this approach. According to the statisticians who devised and work within this methodology, when this philosophy is put into practice it allows for the estimation of student ability that is independent of item difficulty. In other words, it does not matter if one version of the test is slightly harder or easier than its predecessor – student ability should still be fairly and comparably estimated.

Of course, items and students rarely fit this model perfectly. Questions whose scoring profiles are glaringly irregular are dutifully removed from the test, and students whose answer sheets are obviously ill-fitting are considered inadequately assessed. However, these misfits are a matter of degree, and human judgment enters into the decision-making process. It is in this way that irregularities in the score distributions can enter. Slightly different sets of anchor items, or minor variations in the order of items can have subtle but appreciable effects on the meaning attached to scores.

None of this is new to the administrators of the FCAT. As the Human Resources Research Organization (HumRRO) said in their audit of the 2010 tests:

> "The factors that create uncertainty act to produce scale scores that are likely to be higher or lower than they might be if equating could be perfect. The issue is not whether equated scores are too high or too low because that is essentially a given. *The issue is with the magnitude of the uncertainty and when 'too high' or 'too low' becomes unacceptably too high or unacceptably too low."* (Exploration of FCAT Equating and Its Impact on Student, School, and District Developmental Scale Scores for 2003 through 2010, Memorandum of Agreement, July 13, 2010.)

When the score irregularities are analyzed with aggregation at the state level, they can appear to be minor and within historic trends. However, what looks like a "bump in the road" at the state level can feel more like "falling off a cliff" at the school level. Depending on where in the cutoffs for achievement levels the score distribution changes occur, small average differences can amass, resulting in numerous schools dropping a grade level. Even if, in the scoring and equating of the

tests, no one has actually committed a mistake and no one can be held blameworthy, radical and pervasive school grade incongruities cannot be lightly tolerated.

Recommendations

As long as tests designed for assessing student ability are used for school accountability, there will be inequities and injustices. It is inherent in the imprecision with which we measure achievement. Nonetheless, it seems an inescapable and important element in helping to identify effective teaching. Our trust in the aggregate measures derived from the FCAT, in its current and future forms, can be enhanced in a few practical ways.

Convergent validity. Prior to 2008, all students in grade 3 through 10 were also required to take the FCAT Reading and Mathematics Norm-Referenced Test (NRT). Besides being useful in individual student progression plans, it provided an invaluable adjunct to the regular FCAT Sunshine State Standards (SSS) tests. The student scores reported in terms of national percentiles provided the students, teachers, and parents with an easily understandable interpretation of their current abilities and progress. Moreover, as a crosscheck for the SSS scores, it provided a useful form of confirmation that went a long way toward quelling any validity concerns. If the NRT tests were reinstated strictly for the purposes of providing comparison data and confirmation to the SSS tests, only very small samples would be necessary, keeping the costs to a minimum.

Spreading over time. If the score components for grading schools were not the yearly results, but the average of the most recent few years, the unreliability issues would be largely subdued. It would be likely that random measurement error in one direction one year would be counterbalanced by the random error from other testing occasions. It would still be possible and desirable to report yearly performance scores that would be sensitive to special programs and intensive efforts while providing early warnings of potential shortfalls. With averaging over time, the application of consequences, in both rewards and punishments, for observable achievement score changes would depend upon consistent, reliable trends.

Supplemental data sources. When it comes to evaluating individual teachers, inferring teacher effectiveness from the single source of student test scores is inadequate. Other measures, including structured observations of teachers, expert review of portfolios of teachers' lesson plans and other materials, and assessments of teacher competencies, knowledge, and skills, would be essential ingredients. Most importantly, these other methods provide the supplemental information necessary to illuminate the ways in which a teacher may improve in order to be more effective.

In the next few years, the State is planning several substantial modifications to the FCAT testing program. New standards will be initiated at all grade levels, end-of-course exams in several content areas will be introduced, and computer-based administration of tests will be gradually implemented. It will be a time of great change and rapid adaptation during which it will be very difficult to interpret meaningfully changes in academic achievement between the previous and new conditions. However, this period will also present a unique opportunity to introduce other kinds of adjustments into the school grading aspects of the testing plan that can have a substantial effect on the perceived trustworthiness of the accountability program.