



THE EDUCATIONAL  
**Forum**

# Learning to Take Tests or Learning for Understanding? Teachers' Beliefs about Test-Based Accountability

by Brett D. Jones and Robert J. Egley

## *Abstract*

*The results of a survey on teachers' perceptions regarding Florida's test-based accountability program raised serious doubts about whether testing has precipitated positive outcomes in upper-elementary students' learning. Nearly all of the 708 Florida upper-elementary teachers who completed the survey reported that testing had a negative effect or no effect on student learning in reading, writing, and mathematics. Factors associated with students' decrease in learning are discussed and implications are provided.*

Recent state and national legislation, such as the No Child Left Behind Act of 2001, attempts to improve education through the use of statewide standards and test-based accountability. Policy makers and testing proponents claim that such test-based accountability programs hold educators and students accountable and, thus, raise student achievement (Evers and Walberg 2002; Raymond and Hanushek 2003). As a result, testing programs have been implemented across the nation to measure student achievement and school quality.

Some educators, researchers, parents, students, and national educational organizations are not convinced that testing programs are the best means to ensure that students are learning and that teachers are teaching effectively. In fact, several researchers have found that teachers often are opposed to testing because of the many negative consequences associated with it (Urdu and Paris 1994; Jones et al. 1999). Though many unintended consequences of testing have been noted (Jones, Jones, and Hargrove 2003), one of the most often cited is the negative effect that it has on teaching practices and, consequently, student learning.

Despite all the positive and negative consequences of testing that have been discussed over the past few years, one of the most important questions regarding testing

remains largely unanswered: Does test-based accountability result in increased student learning? One way to answer this question is to examine student achievement on state-wide tests. Though this might sound like an irrefutable means to answer this question, test scores increase and decrease due to a variety of reasons not necessarily related to increased student learning, such as test score pollution resulting from teaching test-taking skills, changes in the test format and questions, and scoring from year to year. As Linn (2000, 7) stated, "Common sense and a great deal of hard evidence indicate that focused teaching to the test encouraged by accountability uses of results produces inflated notions of achievement when results are judged by comparison to national norms." This type of evidence was provided by Amrein and Berliner (2002) who found that student learning generally stayed the same or decreased after high-stakes tests were implemented. Results such as these indicate that achievement as measured by test scores does not necessarily reflect gains in student understanding.

*Current learning research emphasizes the importance of teaching for understanding, which occurs as facts are connected and organized around concepts.*

To study the effects of test-based accountability on student learning, the authors chose a different strategy: to ask teachers who work with students on a daily basis and who know best how their teaching practices are affected by their state's testing program. Upper-elementary teachers in Florida were surveyed about how that state's testing program had affected their teaching practices and students' learning. Specifically, the teachers were asked:

- how testing had influenced their ability to use effective teaching methods;
- how professional development had helped them with their overall teaching performance and students' performance on tests;
- how much time their students spent practicing test-taking strategies; and
- how useful and accurate the test results were for assessing students' strengths and weaknesses.

This paper describes the results of this study and provides recommendations based on these results.

### ***Background and Literature Review***

**Pressure and Support.** One way to affect change in schools is to provide the right mix of pressure and support (Fullan 1991; Olson 2001). In recent years, policy makers have used test-based accountability to provide the pressure necessary to force change. The logic behind such a policy is that testing will pressure students into working harder and teachers into teaching better. Researchers have provided evidence that testing can affect the amount of pressure that teachers feel (Firestone and Mayrowetz 2000; Pedulla et al. 2003). For example, Jones et al. (1999) found that 88.9 percent of North Carolina teachers reported that their jobs were more stressful since the implementation of the high-stakes testing program.

Some researchers suggested that the pressure of high-stakes testing affects teaching practices differently—for some teachers a positive effect, while for others a negative or little to no effect (Cimbricz 2002; Jones et al. 2003). Others have found that pressure through testing has more of an effect on the content taught than on teaching practices (Firestone and Mayrowetz 2000). Schorr and Firestone (2001) found that while testing did have an effect on teaching practices, changes in methods often were cosmetic rather than deep.

Changes within schools also are dependent on the amount of support teachers receive. Several types of supports are available to teachers, including professional development, in-school trainers, other teachers, materials, and curricular assistance (Firestone, Monfils, and Schorr 2002). The role of support within the context of a testing environment has been studied by Firestone, Monfils, and Camilli (2001, 31) who concluded, “While pressure is effective in getting teachers to attend to new standards, several types of support will be needed to help teachers adopt the instructional strategies associated with the highest level of national standards.” Clearly, support plays an important role in helping teachers to improve their teaching effectiveness.

**Student Learning.** To judge how testing has affected student learning, what it means to “learn” must be considered. Current learning theories emphasize the importance of understanding, as opposed to rote memorization of facts (National Research Council [NCR] 1999). Though learning facts about a subject is important, this should not be the sole focus of learning. Instead, experts in any field of study organize their problem solving around big and important concepts (Voss et al. 1983). The NRC (1999, 9) stated:

*The new science of learning does not deny that facts are important for thinking and problem solving. . . . However, the research also shows clearly that ‘usable knowledge’ is not the same as a mere list of disconnected facts. Experts’ knowledge is connected and organized around important concepts (e.g., Newton’s second law of motion); it is ‘conditionalized’ to specify the contexts in which it is applicable; it supports understanding and transfer (to other contexts) rather than only the ability to remember.*

*Test-based accountability policies should focus less on pressuring educators into compliance and more on providing support through quality professional development.*

Major investments of time are needed to develop understanding and expertise in an area. This point is well stated by the NRC (1999, 24):

*Learning with understanding is often harder to accomplish than simply memorizing, and it takes more time. Many curricula fail to support learning with understanding because they present*

*too many disconnected facts in too short a time—the ‘mile wide, inch deep’ problem. Tests often reinforce memorizing rather than understanding.*

Current learning research emphasizes the importance of teaching for understanding, which occurs as facts are connected and organized around concepts. Moreover, it takes time to reach a level of deep understanding in a subject and to develop the ability to transfer this understanding to different contexts.

### ***Florida Comprehensive Assessment Tests***

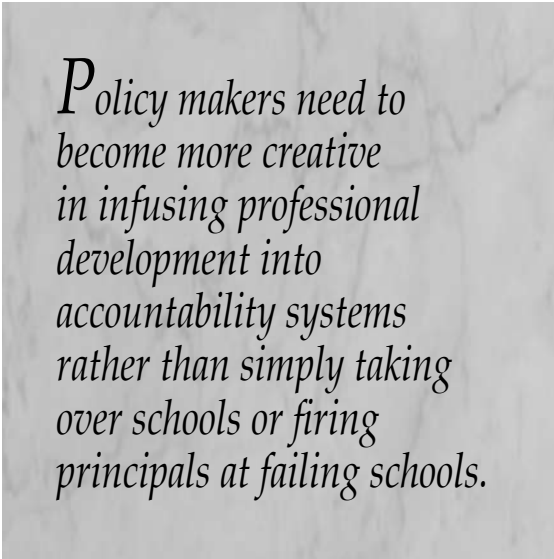
The centerpiece of Florida's test-based accountability program—the Florida Comprehensive Assessment Tests (FCATs)—were first administered in Florida's public schools and used for accountability purposes in spring 1999 (Florida Department of Education 2005). Since that time, schools have been assigned a rating ranging from "A" (making excellent progress) to "F" (failing to make adequate progress), and school grades have been directly linked to accountability rewards and sanctions (Florida Department of Education 2005). The FCAT is considered a strong accountability system because the scores are used as a basis for distributing rewards to higher-performing schools. This contrasts with what Firestone et al. (2001) referred to as weak accountability systems where the pressure of the test scores themselves (with no rewards) are expected to encourage lower-performing schools to improve.

During the year of this study, the FCAT consisted of a criterion-referenced test that measured state standards in reading, writing, and mathematics and a norm-referenced test that measured student performance against national norms. The reading and mathematics tests were administered in grades 3–10, and the writing test

was administered in grades 4, 8, and 10. The FCAT consisted of multiple-choice items at all grade levels tested and "performance items" (requiring a written answer) in reading in grades 4, 8, and 10 and in mathematics in grades 5, 8, and 10. Test results were provided at the student, school, district, and state level. More information regarding the FCAT is available online at [www.firn.edu/doe/sas/fcat.htm](http://www.firn.edu/doe/sas/fcat.htm).

### ***Method***

**Participants and procedure.** Third-, fourth-, and fifth-grade teachers across Florida were surveyed. All 67 Florida school districts were invited to participate in this study; 34 districts (50.7 percent of all districts) agreed to participate. Principals of all the participating elementary schools were contacted three times: twice by electronic mail (e-mail) and once by letter. In the e-mail correspondence, principals were asked to tell their teachers about the survey and to provide them with the Web site address for the online survey. In the letter correspondence, copies of a one-page flyer were included that explained the study and provided the Web site address for the online survey. Principals were asked to distribute the flyers to their third-, fourth-, and fifth-grade teachers. Though the number



*Policy makers need to become more creative in infusing professional development into accountability systems rather than simply taking over schools or firing principals at failing schools.*

of principals that asked their teachers to participate is unknown, completed surveys were received from 708 third-, fourth-, and fifth-grade teachers from 30 school districts (45 percent of all districts) in Florida. The participating teachers represented 235 of the 631 schools in the participating districts (37.2 percent of the schools were represented). The percentage of teachers and schools participating in this study was similar to the statewide percentage of elementary schools at each school grade level.

Most of the teachers were female (88.5 percent) and European American (91.0 percent), while 5.3 percent were African American, 2.6 percent were Hispanic, and 1.1 percent were of another race or ethnicity. Teachers ranged in age from 22 to 68 years old ( $M=41.2$  years old) and had taught school from 1 to 45 years ( $M=13.4$  years). Just over one-fourth (25.2 percent) of the teachers taught third grade, 37.4 percent taught fourth grade, 28.9 percent taught fifth grade, and 8.5 percent taught in a multiage classroom with at least some students in the third, fourth, or fifth grade.

**Survey Instrument.** Teachers completed an anonymous, online questionnaire that queried them about their demographic information, their current teaching practices, and their beliefs about the FCAT. Of the non-demographic informational items reported in this study, 23 items required teachers to respond on a Likert scale, one item required a “yes” or “no” response, and one open-ended item required a written response.

**Data Analysis.** Descriptive statistics were computed for all the closed-ended items. T-tests and ANOVAs were completed for some of the items to test for differences between groups. One-sample t-tests were conducted for three of the Likert-scale items to assess whether the FCAT had an effect—either positive or negative—on students’ learning or on teaching practices. The scale value of “4” was selected as the comparative mean for the one-sample t-tests because it indicated that there had been no change due to the FCAT. That is, if the FCAT had no effect on a teacher, he or she would have selected the value of 4 (labeled as “the same,” “does not influence me,” or “no effect”).

The open-ended item asked teachers to explain why they thought the FCAT program was taking Florida’s public schools in the right or wrong direction. For this item, the overall analysis strategy involved a microanalysis of the teachers’ responses based on a grounded theory approach to qualitative data (Strauss and Corbin 1998). Three researchers conducted this analysis, which resulted in 64 coding categories. Only the results of the categories relevant to the discussion in this paper are presented. A complete description of the coding procedure and final categories can be found in “Voices from the Frontlines: Teachers’ Perceptions of High-Stakes Testing” (Jones and Egley 2004).

## ***Results and Discussion***

The purpose of this study was to examine teachers’ perceptions of how Florida’s testing program had affected upper-elementary students’ learning. As one measure of teachers’ perceptions of student learning, teachers were asked (on a 7-point Likert scale ranging from 1=much less to 7=much more): “If students didn’t have to take the FCAT, your students’ level of knowledge and skill at the end of the year would be:” (Table 1). The mean value for reading was 4.92 ( $SD=1.15$ ); for writing, 4.59 ( $SD=1.20$ ); and for math-

ematics, 4.86 (SD=1.16). One-sample t-tests were conducted to test the null hypothesis that the mean values were equal to a Likert-scale value of “4” (the same). The mean value was significantly different from 4 in reading (t=21.45, p<.001, d=0.80), writing (t=13.04, p<.001, d=0.49), and mathematics (t=19.69, p<.001, d=0.74).

**Table 1. Percentage of Teachers for Each Questionnaire Item**

Questionnaire Item	Subject	% of Teachers Selecting Each Value on a 7-point Likert Scale						
		1	2	3	4	5	6	7
If students didn't have to take the FCAT, your students' level of knowledge and skill at the end of the year would be: <sup>a</sup>	Reading	0.3	0.7	1.8	44.5	21.0	19.3	12.3
	Writing	0.6	2.4	7.0	49.4	18.4	12.1	10.0
	Math	0.4	1.0	2.4	45.1	22.5	16.1	12.4
How does the FCAT influence your ability to use what you consider to be effective teaching methods in: <sup>b</sup>	Reading	11.7	9.5	14.2	35.6	15.1	8.8	5.1
	Writing	10.9	8.6	10.0	35.5	18.9	9.0	7.0
	Math	9.3	10.2	13.9	32.4	17.4	10.5	6.3
How much has the professional development you have received at this school helped improve your overall teaching performance? <sup>c</sup>		3.8	7.2	6.7	31.4	19.2	17.2	14.5
How much has the professional development you received at this school helped improve your students' performance on the FCAT? <sup>d</sup>		8.5	9.0	13.2	39.0	17.0	8.2	5.2
What type of effect does the FCAT have on developmentally appropriate practices? <sup>e</sup>	Reading	17.7	14.6	19.5	19.9	19.2	7.1	2.0
	Writing	17.8	13.1	17.1	20.0	21.9	6.7	3.4
	Math	17.2	12.5	18.6	20.8	20.6	8.0	2.3
How much pressure do you feel to improve your students' FCAT scores this year? <sup>f</sup>		1.4	0.4	1.4	11.7	8.4	14.7	61.8
How useful are the FCAT results for helping you assess students' strengths and weaknesses in: <sup>g</sup>	Reading	19.4	13.4	14.5	37.3	8.6	4.5	2.3
	Writing	22.3	13.7	13.3	36.3	8.1	3.8	2.6
	Math	18.4	12.3	15.2	35.9	9.2	6.3	2.7
How accurate is the FCAT in assessing students' knowledge and skills in: <sup>h</sup>	Reading	10.0	15.3	16.9	40.8	11.5	4.2	1.3
	Writing	10.1	15.9	18.9	37.5	10.0	5.9	1.7
	Math	8.6	16.0	17.9	37.4	13.9	4.4	1.7

<sup>a</sup> 1=much less, 4=the same, and 7=much more

<sup>b</sup> 1=negatively influences me, 4=does not influence me, and 7=positively influences me

<sup>c</sup> 1=has not helped me at all, 4=has helped me to some degree, and 7=has helped me a lot

<sup>d</sup> 1=has not helped them at all, 4=has helped them to some degree, and 7=has helped them a lot

<sup>e</sup> 1=negative effect, 4=no effect, and 7=positive effect

<sup>f</sup> 1=no pressure, 4=some pressure, and 7=a lot of pressure

<sup>g</sup> 1=not useful at all, 4=useful to some degree, and 7=very useful

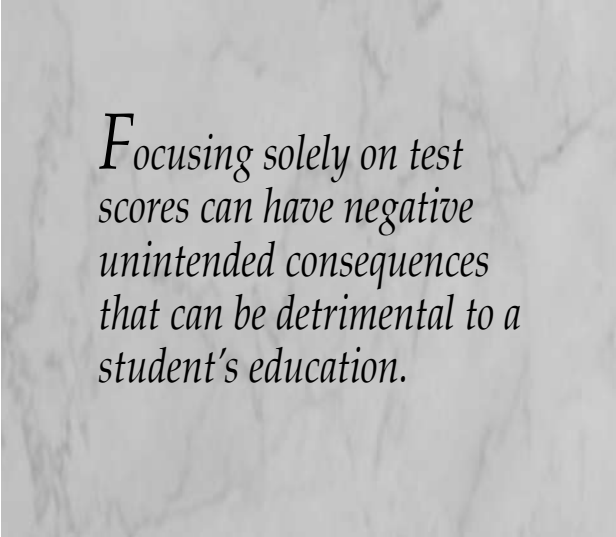
<sup>h</sup> 1=not accurate at all, 4=accurate to some degree, and 7=very accurate

These results indicated that nearly all teachers found the testing program to impede students' learning or to have no effect on it. This finding also was evident in the open-ended item that asked teachers to explain whether they thought the FCAT program was taking public schools in the right direction. More than a third of teachers (35.2 percent) responded that the testing had negative effects on teaching and learning, with 6.2 percent specifically citing that the test takes time and focus away from learning. Other responses cited by teachers were that testing:

- did not accurately measure learning and development (15.7 percent);
- forced teaching that was not developmentally appropriate (3.9 percent);
- stifled student creativity or pushed students into a mold (3.6 percent); and
- did not allow teachers to meet the learning needs of students (2.8 percent).

These beliefs helped explain why nearly half of the teachers reported that students would have more knowledge and skill at the end of the year if they did not have to take the FCAT. More than 90 percent believed that students would learn the same amount or more without the tests. This finding is in direct contrast with the intent of Florida's

testing program, which was designed to "increase student achievement by implementing higher standards" (Florida Department of Education 2005). Clearly, there is a mismatch between the stated intent of Florida's testing policy and the perceptions of many teachers.



*Focusing solely on test scores can have negative unintended consequences that can be detrimental to a student's education.*

In the following sections, factors are presented that explain why most teachers in the survey believed that testing has not had a positive effect on student learning. First, testing has a negative impact on some teachers' ability to use effective teaching methods, especially developmentally appropriate practices. Second, students spend a

great deal of time practicing test-taking strategies, which takes time away from learning and increasing content knowledge. Third, the test results are not perceived as being very accurate or useful for assessing students' strengths and weaknesses; therefore, the test results likely have little effect on changing the teaching and learning processes.

### ***Effects on Teaching Practices***

Teachers were evenly divided (Table 1) when asked: "How does the FCAT influence your ability to use what you consider to be effective teaching methods?" The mean value was 3.80 (SD=1.58) for reading, 3.98 (SD=1.62) for writing, and 3.95 (SD=1.60) for mathematics. One-sample t-tests were conducted to test the null hypothesis that the mean values were equal to the Likert-scale value of "4" (does not influence me). The mean value was significantly different from 4 in reading ( $t=-3.39$ ,  $p=.001$ ,  $d=-0.13$ ), but not in writ-

ing ( $t=-0.31$ ,  $p=.761$ ,  $d=-0.01$ ) or mathematics ( $t=-0.83$ ,  $p=.407$ ,  $d=-.03$ ). These findings indicated that, on average, teachers believed that the FCAT negatively influenced their ability to use effective teaching methods in reading, but had no effect on their ability to use effective teaching methods in writing or mathematics. These results are consistent with findings on how high-stakes testing influenced teachers' teaching practices in other states such as New Jersey and North Carolina (Cimbricz 2002; Firestone et al. 2002; Jones et al. 2003).

Taken at face value, these findings suggested that the testing policy had no overall positive or negative effect on teaching practices. However, the results did not provide an indication of the extent of changes in methods by the teachers. Future research should examine whether the changes were significant or merely cosmetic, as reported by researchers in New Jersey (Schorr and Firestone 2001).

**Professional Development**

The authors also were interested in how professional development affected teachers' instruction. When asked specifically, "How much has the professional development you have received at this school helped improve your overall teaching performance?" teachers reported that it had helped them to some degree ( $M=4.64$ ,  $SD=1.57$ ). Similarly, when asked how much professional development helped improve their students' performance on the FCAT, they reported that it helped them to some degree ( $M=3.92$ ,  $SD=1.49$ ).

The effects of professional development varied according to how teachers perceived that the FCAT affected their teaching methods. Teachers who perceived the FCAT to have a positive influence on their teaching methods reported that professional development had helped their overall teaching performance and students' performance, compared to those teachers who perceived that the FCAT had a negative influence on their methods (Table 2).

**Table 2. Mean Comparisons by Teachers Who Perceived the FCAT to Have a Positive, Negative, or No Influence on Their Ability to Use Effective Teaching Methods**

Dependent Variable	Negative Influence (n = 264)	No Influence (n = 170)	Positive Influence (n = 253)	F-Value
Effect of professional development on teachers' overall teaching performance <sup>1</sup>	4.33 <sup>c</sup>	4.60 <sup>c</sup>	5.01 <sup>a,b</sup>	12.33*
Effect of professional development on students' performance on FCAT <sup>2</sup>	3.41 <sup>b,c</sup>	3.76 <sup>a,c</sup>	4.57 <sup>a,b</sup>	45.29*

\*  $p<.001$

<sup>1</sup> Reported on a 7-point Likert scale: 1=has not helped me at all; 4=has helped me to some degree; and 7=has helped me a lot.

<sup>2</sup> Reported on a 7-point Likert scale: 1=has not helped them at all; 4=has helped them to some degree; and 7=has helped them a lot.

Scheffe mean comparisons were used to test all possible pairs. Different superscripts for a particular variable indicate differences between groups at the  $p<.05$  level. Superscript "a" indicates the negative influence group, "b" indicates the no influence group, and "c" indicates the positive influence group.



Though these results are correlative, they are consistent with the findings of other researchers who discovered that teachers who receive professional development support are more likely to make positive changes in their teaching (Firestone et al. 2001; Desimone et al. 2002). All suggested that given the same amount of pressure from the state's testing program, teachers with more professional development support are more likely than teachers with less support to perceive the FCAT to positively impact their ability to use effective teaching methods. The implication is that teachers' instruction may benefit from quality professional development. Further studies should be conducted to determine the type of professional development that teachers find most helpful and whether or not these types of professional development are aimed at increasing student understanding

or only at increasing test scores on standardized tests.

*Teachers in the classroom are in a better position to assess student learning through student portfolios and other measures tied directly to specific course objectives.*

### ***Developmentally Appropriate Practices***

Teachers' responses generally were negative when asked about the FCAT's effect on developmentally appropriate practices. Nearly half of the teachers (Table 1) reported that the testing had a negative effect when asked: "What type of effect does the FCAT have on developmentally appropriate practices?" The mean value in reading was 3.38 (SD=1.63); in writing was 3.49 (SD=1.68); and in mathematics was 3.49 (SD=1.64). One-sample t-tests were conducted to test

the null hypothesis that the mean values were equal to the Likert-scale value of "4" (no effect). The mean value was significantly different from 4 in reading ( $t=-10.11$ ,  $p<.001$ ,  $d=-0.38$ ), in writing ( $t=-8.00$ ,  $p<.001$ ,  $d=-0.30$ ), and in mathematics ( $t=-8.21$ ,  $p<.001$ ,  $d=-0.31$ ), indicating that, on average, teachers believed that the FCAT had a negative effect on developmentally appropriate practices in these subjects.

In examining their responses to this open-ended question, the researchers found that teachers had these negative perceptions because the tests did not consider children's different developmental rates. For example, all third graders were expected to take the same test on the same day of the year, regardless of their developmental level. One teacher explained:

*The FCAT focuses on too difficult of concepts for many third graders—and it makes children feel like they are failures in math and they're only in the third grade! Many concepts that we are now expected to teach (like decimals) are very difficult for children because they are not developmentally appropriate. I just taught my class a whole unit on decimals and they could pass the final test, but they didn't really understand that a decimal is less than one! They shouldn't have to—they are only 8–9 years old! They are not*

*developed enough with their abstract thinking to truly understand some math concepts that the FCAT tests. I can teach them to jump through the hoops to pass the test, but true understanding is not happening—and it's really demotivating to me as a teacher.*

Forcing students to learn concepts that are beyond their reach was seen as developmentally inappropriate. Though this teacher taught students to take and pass the test, she did not consider this type of teaching effective in fostering learning for understanding. The obvious recommendation is to ensure that tests are developmentally appropriate for the students taking them.

### ***Teaching Test-Taking Strategies***

Teachers were asked: “In an average week, what percent of instructional time do your students spend practicing test-taking strategies specifically designed to help them score higher on the FCAT? (Only consider the strategies that they wouldn’t practice if the FCAT was not given; assume that the Sunshine State Standards would still be in place.)” Teachers reported spending an average of 43.0 percent of their mathematics instructional time, 42.6 percent of their writing instructional time, and 38.0 percent of their reading instructional time on test-taking strategies. This is an enormous amount of time considering the question specifically instructed teachers to consider only strategies that students wouldn’t practice if the FCAT was not given. The fact that teachers spent so much of their time practicing for the tests helps to explain why some teachers believed that the tests had negatively affected students’ learning. Without the tests, teachers could use this time to teach students knowledge and skills in reading, writing, and mathematics or in other non-tested subjects such as science, social studies, music, and art.

When asked to explain whether they thought that the FCAT program was taking public schools in the right direction, 23.3 percent of the teachers reported that the testing forced teaching to the test and test preparation. One teacher described how teaching to the test is not the same as “real learning” (i.e., learning for understanding):

*Schools aren't improving their academics as students score better on the FCAT. They are just taking more time to teach to the test and, unfortunately, away from real learning. We aren't getting smarter students; we are getting smarter test takers. That is NOT what we are here for! The schools that score well are focusing on teaching to the test at a very high cost to their students.*

Similar to other states with strong accountability systems (e.g., North Carolina), Florida teachers reported that they felt a lot of pressure to im-

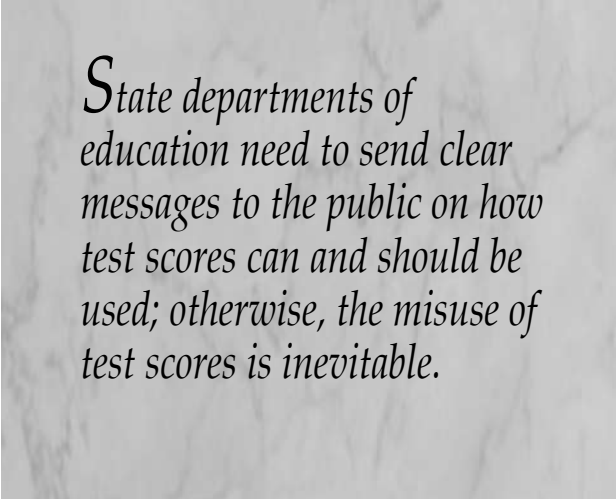
*Rather than providing useful data at the student level, high-stakes tests might be more useful on a global level to inform teachers and administrators about school-wide and district-wide trends.*

prove students' test scores ( $M=6.17$ ,  $SD=1.3$ ). Nearly all teachers (96.7 percent) reported that the pressure they felt was between "some pressure" and "a lot of pressure" to improve their students' test scores (Table 1). Nearly two-thirds (61.8 percent) of teachers selected the highest value, indicating that they felt "a lot of pressure" to improve students' scores. This finding suggested that the testing program provided the pressure that some policy makers intended.

Unfortunately, this pressure is also forcing teachers to spend more time teaching test-taking strategies. One teacher reported in the open-ended question: "Teachers I know, including myself, have simply begun teaching to the test due to the pressure from the administration and the county."

To assess whether the level of pressure teachers felt resulted in more time spent on test-taking strategies, teachers were placed into two groups based on their response to the item about pressure. One group included 535 teachers who felt the greatest pressure

(reported a value of 6 or 7 on the Likert scale). The other group included 141 teachers who felt some pressure (reported a value of 4 or 5 on the Likert scale). T-tests were conducted to compare the differences spent on test-taking strategies between these two groups.



*State departments of education need to send clear messages to the public on how test scores can and should be used; otherwise, the misuse of test scores is inevitable.*

Teachers who felt the most pressure reported spending a significantly higher percentage of their instructional time on test-taking strategies in reading, writing, and mathematics. The percentage of time spent teaching test-taking strategies in reading was 40.9 percent for teachers

who felt the most pressure and 28.3 percent for teachers who felt some pressure ( $t=6.05$ ,  $p<.001$ ,  $d=0.54$ ). Similarly, the percentage of time spent teaching test-taking strategies in writing was 45.9 percent for teachers who felt the most pressure and 32.6 percent for teachers who felt some pressure ( $t=5.18$ ,  $p<.001$ ,  $d=0.47$ ). The percentage of time spent teaching test-taking strategies in mathematics was 46.1 percent for teachers who felt the most pressure and 33.6 percent for teachers who felt some pressure ( $t=5.61$ ,  $p<.001$ ,  $d=0.51$ ).

Pedulla et al. (2003) reported that teachers in states with high-stakes tests were more likely than teachers in states with lower-stakes tests to spend a greater amount of time on test preparation. However, the results of the study described in this paper suggested that even within a state, teachers who felt more pressure spent more time in test preparation. The teachers who felt the most pressure in Florida spent an average of 13 percent more of their instructional time teaching test-taking strategies in read-

ing, writing, and mathematics. These findings suggested that the amount of pressure matters, as does how the teachers perceive and interpret the pressure.

Two ANOVAs were conducted to assess whether other factors affected the amount of time spent practicing test-taking strategies. The findings revealed that grade level taught (third, fourth, or fifth) impacted the amount of time teachers spent teaching test-taking strategies in some subjects, but the grade ranking earned by the school—ranging from A to F—did not.

Fourth-grade teachers reported spending more time practicing for the reading test than fifth-grade teachers and more time practicing for the writing test than both the third- and fifth-grade teachers (Table 3). Considering that the writing test is administered at the end of the fourth grade and not at the end of the third or fifth grade, this finding was not unexpected. Rather, it supported the notion that pressure is a factor in the amount of test preparation in which teachers engage. Fourth-grade teachers likely felt more pressure than the third- and fifth-grade teachers to improve students' writing scores.

**Table 3. Mean Comparisons by Grade Level for Percentage of Time Spent Practicing Test-Taking Strategies<sup>1</sup>**

Subject	Third Grade (n=178)	Fourth Grade (n=262)	Fifth Grade (n=201)	F-Value
Reading	38.9%	41.9% <sup>c</sup>	34.1% <sup>b</sup>	5.34*
Writing	39.9% <sup>b,c</sup>	56.7% <sup>a,c</sup>	27.3% <sup>a,b</sup>	67.40**
Math	42.4%	42.7%	45.8%	1.05

\* p<.01; \*\* p<.001

<sup>1</sup> Reported on an 11-point Likert scale ranging from 0 percent to 100 percent.

Scheffe mean comparisons were used to test all possible pairs. Different superscripts for a particular variable indicate differences between groups at the p<.05 level. Superscript "a" indicates the third-grade group, "b" indicates the fourth-grade group, and "c" indicates the fifth-grade group.

No significant differences were evident in the amount of time spent practicing for tests based on the school's grade in reading (F[4, 651]=1.74, p=.14), writing (F[4, 644]=0.95, p=.47), or mathematics (F[4, 649]=0.90, p=.47). For example, "A" schools did not spend significantly more or less time practicing for tests than "C" schools. This finding indicated that the pressure for students to perform was felt by teachers at all schools, regardless of the school's grade, not just at the lower-performing ones.

*The teachers in this study who spent the most time teaching test-taking strategies were those who reported feeling the most pressure.*

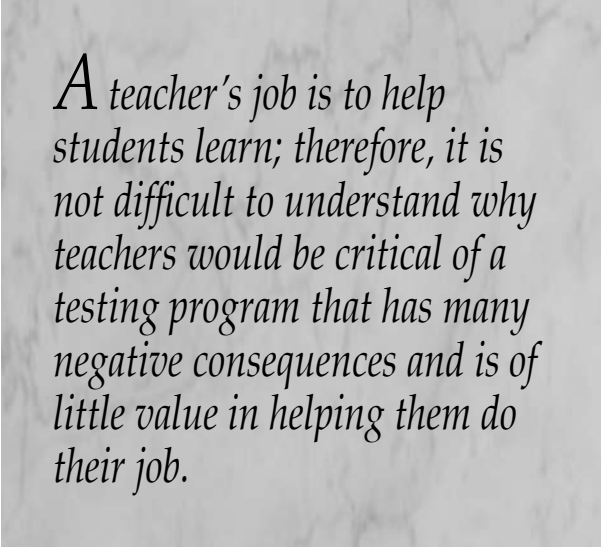
The implications of the findings in this section are that policy makers need to find ways to minimize the amount of pressure felt by teachers to improve students' test scores. Doing so would likely reduce the amount of time teachers spend teaching test-taking strategies. The teachers in this study who spent the most time teaching test-taking strategies were those who reported feeling the most pressure. Reducing the amount of time spent on test-taking strategies would allow more time to teach for understanding.

### ***Usefulness and Accuracy of Test Results***

Two recent reports (Commission on Instructionally Supportive Assessment 2001; NRC 2001) indicated that part of the purpose of testing should be to provide teachers with information they can use to improve their instruction. Statewide assessments should support both accountability and instruction. Popham (2000, 80) stated, "If educational assessment doesn't help children learn better, we shouldn't be doing it."

One of the questions in this study was whether the FCAT results provided teachers with information that was useful and accurate in assessing students. Teachers were asked,

"How useful are the FCAT results for helping you assess students' strengths and weaknesses?" Teachers did not find the results to be very useful in reading ( $M=3.25$ ,  $SD=1.53$ ), writing ( $M=3.16$ ,  $SD=1.56$ ), or mathematics ( $M=3.35$ ,  $SD=1.57$ ) (Table 1). Moreover, teachers did not find the FCAT very accurate in assessing students' knowledge and skills (Table 1). The mean response was 3.46 ( $SD=1.34$ ) for reading, 3.46 ( $SD=1.39$ ) for writing, and 3.52 ( $SD=1.36$ ) for mathematics.



*A teacher's job is to help students learn; therefore, it is not difficult to understand why teachers would be critical of a testing program that has many negative consequences and is of little value in helping them do their job.*

Because teachers claimed that the FCAT results were not very useful or accurate in assessing students, the results likely provided little data to improve student learning.

A teacher's job is to help students learn; therefore, it is not difficult to understand why teachers would be critical of a testing program that has many negative consequences and is of little value in helping them do their job. As one teacher noted:

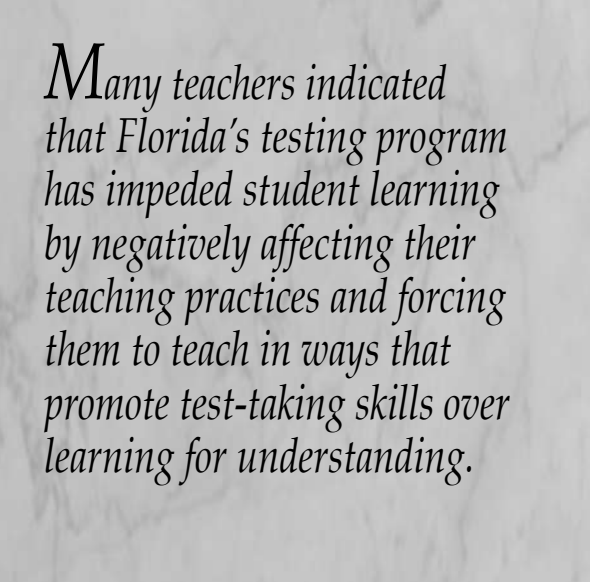
*My personal belief is that the FCAT is a political football and that given the current climate in Tallahassee, its real mission is not to provide accountability to families, communities, etc. or to help schools discern better instructional techniques for students. Rather, the mission is to diminish public education, advance a special interest agenda for charter schools and private education, and advance political careers.*

Clearly, this teacher did not believe that the goal of testing is to improve student learning. Teachers would be more satisfied with the testing program if it met their goal

of improving student learning. To do so, test developers should improve the usefulness of the test results so that teachers can use the results to better assess students' strengths and weaknesses and make positive changes in their teaching practices. The Commission on Instructionally Supportive Assessment (2001) identified nine requirements for large-scale tests to help teachers improve their instruction. Perhaps these suggestions would be helpful to test developers in improving the usefulness of the results to teachers.

### ***Limitations***

As with any research, this study had limitations that may impact its generalizability. First, the results of this study represented the self-reported perceptions of teachers. The actual practices of teachers were not documented. Second, the non-respondents might have had different opinions than those teachers who completed the survey. Third, Florida might have unique factors that make the findings less applicable to testing programs in other states. Finally, the perceptions of the elementary teachers in this study may be different from those of middle or high school teachers. Nonetheless, the findings are believed to provide an accurate picture of how some elementary teachers are reacting to test-based accountability.



*Many teachers indicated that Florida's testing program has impeded student learning by negatively affecting their teaching practices and forcing them to teach in ways that promote test-taking skills over learning for understanding.*

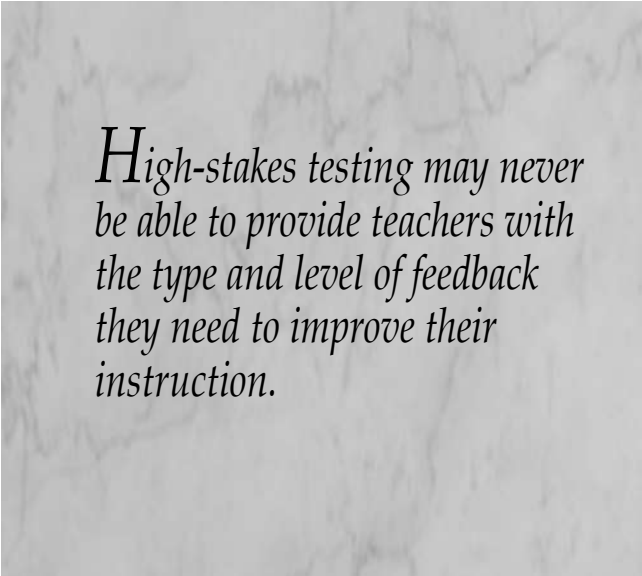
### ***Implications***

This section builds on the implications previously discussed. The results of this study add to a body of research that suggests that test-based accountability by itself does not have a clear overall positive or negative effect on teaching practices. However, an important factor to consider in how teachers perceive the effects of testing on instruction appears to be how they perceive the professional development they receive. Though this study does not provide evidence as to what type or how much support is needed by teachers, teachers who received helpful professional development were more likely to perceive that testing had a positive effect on their ability to use effective teaching methods. As a result, test-based accountability policies should focus less on pressuring educators into compliance and more on providing support through quality professional development. Policy makers need to become more creative in infusing professional development into accountability systems rather than simply taking over schools or firing principals at failing schools. When teachers are provided with quality professional development, teaching practices can improve, even within the context of high-stakes accountability (Jones and Johnston 2004).

Teachers who felt the most pressure from testing reported spending a significantly higher percentage of their time teaching students test-taking strategies. Consequently, reducing the amount of pressure teachers feel to improve students' test scores might reduce the amount of time that teachers spend on test-taking strategies. It is unclear why some teachers feel more pressure than others or in which contexts teachers feel more pressure. However, this study does suggest that testing writing at some grades and not at others pressured teachers at the grades tested to spend more time practicing test-taking strategies.

Administrators might be in a key position to lessen the pressure on teachers by acknowledging it, making efforts to better understand its causes, and providing support that might reduce it. More research is needed to better understand how principals, school district administrators, and state departments of education can support teachers to lessen the pressure to improve test scores and focus more on providing a quality education. The authors are not opposed to teachers increasing test scores, but concerned that focusing

solely on test scores can have negative unintended consequences that can be detrimental to a student's education.



*High-stakes testing may never be able to provide teachers with the type and level of feedback they need to improve their instruction.*

More research is needed to determine the types of feedback from high-stakes tests that are most useful to teachers in helping them improve their instruction. Teachers in this study did not find the test results overly useful or accurate in assessing students' strengths and weaknesses in reading, writing, or mathematics. Yet, high-stakes testing may never be able to provide teachers with the type and level of feedback they need to improve their instruction. One reason is that the

tests are standardized at the state level and, therefore, cannot include the entire curriculum that is taught within any one classroom. Teachers in the classroom are in a better position to assess student learning through student portfolios and other measures tied directly to specific course objectives. Another reason that high-stakes tests are not overly useful to teachers is that standardized tests are costly to implement and score. Therefore, they generally are given only once during the school year. Though some states use practice or "benchmark" tests during the year to assess student progress, such tests take away valuable instructional time and money. Teachers in one Florida school district said that the practice tests are "cumbersome, of little value, and not taken seriously by students" (Tobin 2006, B1). When tests are given once a year, the results usually are received near the end of the school year or in the summer when they are of little or no use to teachers. In addition, students and teachers in Florida have not been allowed to obtain copies of

specific test forms because of the expense associated with creating new tests each year (Associated Press 2003). Because the same test questions are used in subsequent years, allowing students and teachers to examine them could undermine the integrity of future test administrations. As a result, the test items never can be used as tools to help students learn from their mistakes.

Rather than providing useful data at the student level, high-stakes tests might be more useful on a global level to inform teachers and administrators about school-wide and district-wide trends. Whether these types of statewide tests provide valid test scores for individual students is questionable. In its position statement on student accountability standards and high-stakes testing, the North Carolina School Psychology Association (Armistead, Armistead, and Breckheimer 2001, i) stated, "The Students Accountability Standards' use of [high-stakes] test results to make major decisions about individual students is not adequately validated and will cause serious harm to North Carolina's most vulnerable students. The [high-stakes tests were] not developed for making important decisions about individual students." Such statements remind us that we must continually seek to monitor how test scores are being used and for what purposes. State departments of education need to send clear messages to the public on how test scores can and should be used; otherwise, the misuse of test scores is inevitable.

### Conclusion

The results of this study raise serious questions about whether the pressure of test-based accountability has had a positive effect on student learning. Instead of improving teaching and learning, many teachers indicated that Florida's testing program has impeded student learning by negatively affecting their teaching practices and forcing them to teach in ways that promote test-taking skills over learning for understanding. These teachers struggle with the problem of wanting to teach for understanding, yet feeling limited in their ability to do so within the context of a high-stakes environment that measures achievement only through standardized test scores. Policy makers need to seriously consider remedies to reduce these negative effects on learning to ensure that students are taught in ways consistent with current teaching and learning theories and that students receive an education that emphasizes understanding, not simply test-taking skills.

### References

- Amrein, A. L., and D. C. Berliner. 2002. High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives* 10(18). Available at: <http://epaa.asu.edu/epaa/v10n18>.
- Armistead, L., R. Armistead, and S. Breckheimer. 2001. *Grading the North Carolina student accountability standards*. Charlotte, NC: North Carolina School Psychology Association.
- Associated Press. 2003. Anti-FCAT group wants state to release tests. *St. Petersburg Times*, August 12.
- Cimbricz, S. 2002. State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives* 10(2). Available at: <http://epaa.asu.edu/epaa/v10n2.html>.
- Commission on Instructionally Supportive Assessment. 2001. *Building tests to support instruction and accountability: A guide for policymakers*. Washington, DC: National Education Association. Available at: [www.nea.org/accountability/buildingtests.html](http://www.nea.org/accountability/buildingtests.html).
- Desimone, L. M., A. C. Porter, M. S. Garet, K. S. Yoon, and B. F. Birman. 2002. Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis* 24(2): 81-112.
- Evers, W. M., and H. J. Walberg, eds. 2002. *School accountability*. Stanford, CA: Hoover Institution Press.
- Firestone, W. A., and D. Mayrowetz. 2000. Rethinking 'high stakes': Lessons from the United States and England and Wales. *Teachers College Record* 102(4): 724-49.
- Firestone, W. A., L. Monfils, and G. Camilli. 2001. *Pressure, support, and instructional change in the context of a state testing program*. Paper presented at the Annual Meeting of the American Educational Research Association, April 10-14, Seattle, WA.
- Firestone, W. A., L. Monfils, and R. Y. Schorr. 2002. *Test preparation in New Jersey: Inquiry-oriented and didactic responses*. Paper presented at the Annual Meeting of the British Educational Research Association, September 13, Exeter, England.



## Jones and Egley

- Florida Department of Education. 2005. *Florida Comprehensive Assessment Test (FCAT)*. Tallahassee, FL: FLDE. Available at: [www.firn.edu/doe/sas/fcats.htm](http://www.firn.edu/doe/sas/fcats.htm).
- Fullan, M. 1991. *The new meaning of educational change*. New York: Teachers College Press.
- Jones, B. D., and R. J. Egley. 2004. Voices from the frontlines: Teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives* 12(39). Available at: <http://epaa.asu.edu/epaa/v12n39>.
- Jones, B. D., and A. F. Johnston. 2004. High-stakes testing in elementary school: Teachers' perceptions of the effects on teaching and student outcomes. *Research in the Schools* 11(2): 1–16.
- Jones, M. G., B. D. Jones, and T. Hargrove. 2003. *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield.
- Jones, G. M., B. D. Jones, B. Hardin, L. Chapman, T. Yarbrough, and M. Davis. 1999. The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan* 81(3): 199–203.
- Linn, R. 2000. Assessments and accountability. *Educational Researcher* 29(2): 4–16.
- National Research Council. 1999. *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- National Research Council. 2001. *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Olson, L. 2001. Finding the right mix. *Education Week* 20(17): 12–20.
- Pedulla, J. J., L. M. Abrams, G. F. Madaus, M. K. Russell, M. A. Ramos, and J. Miao. 2003. *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston, MA: National Board on Educational Testing and Public Policy, Boston College. Available at: [www.bc.edu/research/nbetpp/statements/nbr2.pdf](http://www.bc.edu/research/nbetpp/statements/nbr2.pdf).
- Popham, W. J. 2000. *Modern educational measurement: Practical guidelines for educational leaders*, 3rd ed. Boston: Allyn and Bacon.
- Raymond, M. E., and E. A. Hanushek. 2003. High-stakes research. *Education Next* 3(3): 48–55.
- Schorr, R. Y., and W. A. Firestone. 2001. *Changing mathematics teaching in response to a state testing program: A fine-grained analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, April 11–14, Seattle, WA.
- Strauss, A., and J. M. Corbin. 1998. *Basics of qualitative research: Techniques and procedures for developing grounded theory*, 2nd ed. Thousand Oaks, CA: Sage.
- Tobin, T. C. 2006. School budget cuts stop at \$19M. *St. Petersburg Times*, March 22.
- Urdan, T. C., and S. G. Paris. 1994. Teachers' perceptions of standardized achievement tests. *Educational Policy* 8(2): 137–56.
- Voss, J. F., T. R. Greene, T. A. Post, and B. C. Penner. 1983. Problem solving skills in the social sciences. In *The psychology of learning and motivation: Advances in research theory*, vol. 17, ed. G. H. Bower, 165–213. New York: Academic Press.



**Brett D. Jones** is an Assistant Professor in the Educational Psychology program within the School of Education at Virginia Tech. Jones studies the effects of test-based accountability programs on teachers and administrators and is coauthor of *The Unintended Consequences of High-Stakes Testing* (Rowman & Littlefield 2003).



**Robert J. Egley** is Principal of Clewiston High School in Hendry County, Florida. He has 18 years' experience as an administrator at the elementary, middle, and high school levels and four years' experience as an Assistant Professor at the University of South Florida–St. Petersburg. Egley's research interests include high-stakes testing, invitational education, and instructional leadership.

