Journal of Applied School Psychology

ISSN: 1537-7903 (Print) 1537-7911 (Online) Journal homepage: http://www.tandfonline.com/loi/wapp20

The Unintended Outcomes of High-Stakes Testing

Brett D. Jones

To cite this article: Brett D. Jones (2007) The Unintended Outcomes of High-Stakes Testing, Journal of Applied School Psychology, 23:2, 65-86, DOI: 10.1300/J370v23n02_05

To link to this article: http://dx.doi.org/10.1300/J370v23n02_05

The Unintended Outcomes of High-Stakes Testing

Brett D. Jones

Virginia Tech

SUMMARY. Although it is important to evaluate the intended outcomes of high-stakes testing, it is also important to evaluate the unintended outcomes, which might be as important or more important than the intended outcomes. The purpose of this paper is to examine some of the unintended outcomes of high-stakes testing, including those related to: (a) using tests as a means to hold educators accountable, (b) the effects on instruction, (c) the effects on student and teacher motivation, and (d) the effects on students who are at-risk of school failure. In examining the evidence, I conclude that while some unintended outcomes of high-stakes testing have been positive, many of the unintended outcomes have been negative. Hopefully, through a greater awareness of the unintended outcomes, school psychologists can work to minimize the negative effects of testing on students and educators. doi:10.1300/J370v23n02_05

Address correspondence to: Brett D. Jones, Virginia Tech, School of Education, 310 War Memorial Hall (0313), Blacksburg, VA 24061 (E-mail: brettjones@vt.edu).

[Haworth co-indexing entry note]: "The Unintended Outcomes of High-Stakes Testing." Jones, Brett D. Co-published simultaneously in *Journal of Applied School Psychology* (The Haworth Press, Inc.) Vol. 23, No. 2, 2007, pp. 65-86; and: *High Stakes Testing: New Challenges and Opportunities for School Psychology* (ed: Louis J. Kruger, and David Shriberg) The Haworth Press, Inc., 2007, pp. 65-86. Single or multiple copies of this article are available for a fee from The Haworth Document Delivery Service [1-800-HAWORTH, 9:00 a.m. - 5:00 p.m. (EST). E-mail address: docdelivery@haworthpress.com].

Available online at http://japps.haworthpress.com

doi:10.1300/J370v23n02_05

KEYWORDS. High stakes tests, unintended outcomes, accountability, instructional effects, motivation

INTRODUCTION

Educators and researchers have noted the unintended outcomes of standardized testing for many years (e.g., Smith, 1991), especially those that have negative effects on students. Recently, however, the widespread use of standardized tests for high-stakes decisions regarding students and educators has magnified the impact of the unintended consequences and created other unforeseen consequences. The purpose of this paper is to examine some of the unintended outcomes of high-stakes testing with a particular focus on those that affect students who are at-risk of school failure. In the first part of this paper, I describe some of the unintended outcomes related to: (a) using tests as a means to hold educators accountable, (b) the effects on instruction, and (c) the effects on student and teacher motivation. In the second part, I discuss the unintended effects on students with learning or behavioral problems, students from economically impoverished families, students from minority groups, and students with limited English proficiency.

Although I refer to the outcomes in this article as *unintended*, it is impossible to determine whether these outcomes were intended or unintended. Therefore, it might be more accurate to define unintended outcomes as those that were not the primary intended outcome according to the *Statement of Purpose* provided in Title I of the *No Child Left Behind Act of 2001* (NCLB, 2002): "The purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments" (Title I, Sect. 1001, 20 USC 6301, para. 1).

UNINTENDED OUTCOMES FOR STUDENTS AND EDUCATORS

The Use of Standardized Tests for Accountability

One of the main purposes of NCLB (2002) was to increase accountability related to student achievement by mandating states to implement statewide assessments (U.S. Department of Education, n.d.). Some

teachers have suggested that such accountability was needed. For example, one Florida teacher reported: "I believe that the [testing] has made teachers accountable for teaching the Sunshine State Standards. We had the Sunshine State Standards, but until there was the accountability, not all teachers were using them" (Jones & Egley, 2004c, *Themes 1 and 6*, para. 14). The viewpoint presented by this teacher suggests that using test scores to hold teachers accountable is working as intended. In fact, 90% of teachers in one study reported that teachers should be accountable for their teaching (Reese, Gordon, & Price, 2004).

Many educators and researchers, however, believe that using high-stakes test scores to hold students, teachers, and schools accountable is improper and unfair (Abrams, Pedulla, & Madaus, 2003; Haney, 2002; Popham, 1999, 2000). Some educators claim that it is unfair to compare students on a one-time standardized test because children develop at different rates and come from different backgrounds. They claim that many factors related to student achievement are out of their control, such as students' parental involvement, socioeconomic status, and home life (Jones & Egley, 2004c). For this reason, they report that it is particularly unfair to compare schools that serve different populations (Jones & Egley, 2004a). Empirical evidence appears to support these concerns. For instance, Reeves (2000) found that 30-40% of the variation in test scores between districts in Kentucky could be attributed to contextual effects that were not under the direct control of teachers and administrators. Similarly, Wilkins (2000) found that nearly 50% of the variance in test passing rates in Virginia was determined by factors unrelated to schooling such as household income and parental education (cited in English, 2002).

Several professional organizations, such as the National Association of School Psychologists (NASP), the American Psychological Association, and the American Educational Research Association (AERA) also support the position that a single test score should not be used to make high-stakes decisions for students. For example, NASP stated that: "NASP strongly opposes the use of large-scale testing as the *sole determinant* for making critical, high stakes decisions about individual students and educational systems, including access to educational opportunity, retention or promotion, graduation or receipt of a diploma" (NASP, 2003, para. 2). AERA made a similar statement: "Decisions that affect individual students' life chances or educational opportunities should not be made on the basis of test scores alone" (2004, para. 6). Clearly, these organizations are opposed to using test scores alone to

make high-stakes decisions; yet, this is exactly how these scores have been used.

One of the problems with relying solely on test scores to make high-stakes decisions is that it involves making inferences about the quality of teachers, administrators, and schools. Popham (2000) explains, from a measurement perspective, why it is unacceptable to make these inferences about educational quality using standardized test scores:

> When standardized achievement tests are employed to ascertain educational quality it really is like measuring temperature with a tablespoon. Tablespoons have a different measurement mission than indicating how hot or cold something is. Standardized achievement tests also have a different measurement mission than indicating how good or bad a school is. Standardized achievement tests should be used to make the comparative interpretations that they were intended to provide. They should not be used to judge educational quality. (p. 400)

Despite these types of warnings by educators and measurement experts, student test scores have consistently been used to rate the quality of schools, especially in states such as Florida where schools are given a letter grade (i.e., A, B, C, D, F) based on students' test scores. In fact, the Governor of Florida, Jeb Bush, has called using test scores to assess schools a "key innovation" (Bush, 2003).

Because the school ratings are reported publicly in newspapers and on state websites (e.g., http://fcat.fldoe.org/), the rating becomes a label for the school. Limiting a school's quality to a rating oversimplifies the complexity of factors that contribute to a quality education. For instance, in Florida, half of a school's grade is based on students' reading test scores, one-third is based on students' mathematics test scores, and one-sixth is based on students' writing test scores. By limiting a school's rating to these academic areas, other important factors are excluded from the rating, such as student work samples, student dropout rate, types and number of courses offered, number of advanced placement courses taken, extracurricular activities available, and students' attitude toward and interest in school (Popham, 2004). As an example of how limiting this view of educational quality is, a parent at one Florida school found the "F" rating of her child's school surprising: "When people look at the grade, they're going to think that the teachers are fail-

ing the students. That has absolutely not been my experience" (Gilmer, 2002, p. B6).

Student test scores are also being used to judge teachers. For example, some school districts (e.g., Denver, Houston) have begun to tie teacher pay to students' test scores, a practice that involves the use of second-level inferences that Popham has warned against (Popham, 2000). In Houston, if students improve on state and national tests, teachers are given as much as $3,000 in extra pay (HISD Connect, n.d.). Teachers who motivate students or help students develop socially or emotionally are not rewarded under such a system. The unintended outcome, therefore, is that the test scores have become the sole measure of teacher and school quality, which severely limits what is considered to be quality teaching and a quality education.

Effects on Instruction

There is strong evidence that high-stakes testing has coerced teachers into aligning their curriculum to the areas tested (e.g., Firestone, Mayrowetz, & Fairman, 1998). On one hand, this may be considered a positive consequence of high-stakes testing in that teachers should be responsible for teaching the state curriculum. As an example, teachers and administrators in one Ohio district found that testing helped the school system align the curriculum between grade levels, helped educators identify curricular weaknesses, and made educators more conscious of educational outcomes (DeBard & Kubow, 2002). Similarly, some teachers in Florida were pleased that the testing had standardized the curriculum across the state and that it had given teachers a standard to which to teach (Jones & Egley, 2004c).

On the other hand, state curricula are too extensive to be accurately measured with a one-time standardized test. As a result, standardized tests are generally limited to only a few subjects such as reading, writing, and mathematics. A possible negative outcome, therefore, is that the curriculum is limited to the subjects tested. Other subjects such as social sciences, health, music, art, and physical education take a back seat and may be excluded completely from the curriculum (Jones, Jones, & Hargrove, 2003).

A related curriculum concern is that the goal of schooling is being restricted to passing standardized tests. School has become limited to developing basic academic and cognitive skills and, in some cases, thinking skills. As a result, the emphasis on other major goals such as developing students' creativity, self-concept, interpersonal relations, ability to be

self-directed, ability to become involved in a democracy, emotional and physical well-being, moral and ethical character, and ability to contribute to the development of a better society (Goodlad, 1979), might be diminishing. For instance, Horn (2003) found that the 10th grade English Language Arts test in Massachusetts was ensuring proficiency in only a subset of skills that have been defined as essential for work in the new millennium.

Some teachers believe that the limited curriculum has made their lessons less engaging for students. Consider this teacher's statement:

> Before [standardized testing] I was a better teacher. I was exposing my children to a wide range of science and social studies experiences. I taught using themes that really immersed the children into learning about a topic using their reading, writing, math, and technology skills. Now I'm basically afraid to NOT teach to the test. I know that the way I was teaching was building a better foundation for my kids as well as a love of learning. Now each year I can't wait until March is over so I can spend the last two and a half months of school teaching the way I want to teach, the way I know students will be excited about. (Jones & Egley, 2004c, *Themes 2 and 7: Effects on the Curriculum*).

A related point made by this teacher is that testing can have a negative effect on students' in-depth learning and understanding. Because some educators believe that the tests cover a wide range of topics in the curriculum areas tested, they might be less likely to devote the time needed for in-depth exploration of a topic. This can be problematic because researchers have found that learning with understanding (as opposed to rote memorization) takes time (National Research Council, 2000). This issue may be worse in states that administer their tests in February and March, a couple of months prior to the end of the school year. In these states, teachers must fit the entire year's worth of curriculum into about two-thirds of the academic year.

In many cases, state standardized testing has not only affected *what* is taught, but also *how* it is taught. Although there does not appear to be any systematic effect of testing on teaching that can be generalized to all teachers and states (Cimbricz, 2002; Jones, Jones, & Hargrove, 2003), several negative effects have been noted. The most commonly cited one on teaching and learning is that teachers feel compelled to teach to the test. Doing so can lead to a focus on low-level knowledge and skills through the use of rote level, discrete, individual drill and skill practice

(Barksdale-Ladd & Thomas, 2000; Hoffman, Assaf, & Paris, 2001). As a teacher in Texas reported: "We try to do hands-on kinds of things actively involving students, but we realize we have to spend lots of time on drill and practice with paper and pencil because of the way the test is formatted" (Gordon & Reese, 1997). In fact, teachers in Florida reported spending an average of 40% of their instructional time practicing test-taking strategies specifically designed to help students score higher on tests (Jones & Egley, 2004b).

The "three-point-five essay" is one example of how Florida's tests have affected instruction. This type of essay gets its name from the fact that a student's response to a prompt on the writing test consists of three points in five paragraphs. Some educators teach this formula to help students pass Florida's 45 minute writing test. The administrator of Florida's Department of Education's Assessment and School Performance Office admits that a three-point-five essay will allow a student to pass the writing test (cited in Catalanello, 2004). The writing project coordinator for one school district in Florida calls it "test-writing"; and unfortunately, she sees it as having negative effects. She states: "We teach the love of writing right out of kids" (cited in Catalanello, 2004). However, with the pressure for students to pass the test, it is understandable why a teacher would choose to use this proven and acceptable method.

Teachers have reported that formulaic approaches have stifled their teaching ability and creativity, including limiting their ability to meet the learning needs of students (Jones & Egley, 2004c). Teachers have noted that students are often not ready for the knowledge and skills they are teaching, but that they have to rush through the curriculum to cover the content before the test. This issue is exacerbated in some districts in states such as North Carolina and Florida that have implemented "pacing calendars" to show which topics should be covered on any particular day. For example, 8,800 third-graders in one Florida school district were scheduled to read *Little Grunt and the Big Egg* from October 13 to October 21 in 2004 (Tobin & Winchester, 2004, October 4). Such a rigid schedule does not allow the flexibility that might be needed to meet the individual needs of students.

Effects on Student and Teacher Motivation

To understand how testing has affected student and teacher motivation, it is helpful to consider some of the various ways in which motivation has been defined and measured in educational settings (see Pintrich & Schunk, 2002, for a complete discussion). One useful definition di-

vides a student's motivation into either intrinsic or extrinsic. Students are intrinsically motivated when they engage in an activity because they enjoy it or are interested in it; whereas, students are extrinsically motivated when they engage in an activity as a means to an end (Pintrich & Schunk, 2002).

High-stakes tests are inherently extrinsic motivators because they focus students on the end result: passing the test. The main reward for passing a high-stakes test is that the student will be allowed to pass to the next grade level and/or that her school will be rated highly (which may result in public praise and monetary rewards for the school). Other rewards have also been given to students who score highly, including limousine rides (George, 2001), new bicycles (George, 2001), and pizza parties (Firestone & Mayrowetz, 2000). These types of rewards make it clear to students that an important aim of schooling is to do well on high-stakes tests (Triplett, Barksdale, & Leftwich, 2003); thus, promoting extrinsic motivation.

Unfortunately, few studies have assessed the impact of testing on student motivation by querying students or by distinguishing between intrinsic and extrinsic motivation. As a result, it is impossible to make definitive statements about how testing has affected students' intrinsic or extrinsic motivation. However, one of the few studies to ask students about their motivation found that 83% of elementary and 45% of secondary students in an Ohio school district agreed that testing had motivated them to study (DeBard & Kubow, 2002). This seems to suggest that at least some students are extrinsically motivated by testing. While this might appear to be a laudable outcome of high-stakes testing, research suggests that extrinsic rewards decrease intrinsic motivation in the long term when perceived as controlling (Deci, 1971; Lepper, Greene, & Nisbett, 1973). The unintended outcome, therefore, is that students might enjoy school subjects less in the future even though they appear to be more motivated in the short term by extrinsic rewards.

In fact, from the inception of the high-stakes testing movement, researchers have warned that high-stakes testing could undermine students' intrinsic motivation (Kellaghan, Madaus, & Raczek, 1996). Researchers who have asked teachers about how testing had affected students' "love of learning" (which is one measure of students' intrinsic motivation) have found that most teachers find testing to have a negative effect on students' love of learning or no effect at all on it (Jones et al., 1999; Rapp, 2002; Yarbrough, 1999). This makes sense given that the focus of testing is on the end result of passing the test with little incentive for teachers to foster students' natural curiosity and intrinsic

motivation. One teacher noted: "School is becoming a drudgery for teachers and students alike. Yes, standards are important and schools should work to ensure every child's success, however, not at the expense of the love of learning" (Jones & Egley, 2004c, *Student Motivation*, para. 2). More research is needed to better understand how testing has affected students' intrinsic motivation and which students are most affected.

There is much evidence, however, to suggest that testing has created a stressful environment for students. Both students and teachers have reported negative effects on students related to testing such as worry, anxiety, nervousness, sweat, tears, stomach aches, irritability, vomiting, headaches, and loss of sleep (DeBard & Kubow, 2002; Hoffman, Assaf, & Paris, 2001; Jones et al., 1999; Triplett, Barksdale, & Leftwich, 2003). In an interesting study, Wheelock, Bebell, and Haney (2000) found that students who were asked to draw a self-portrait in testing situations depicted themselves as anxious, angry, bored, pessimistic, and withdrawn from high-stakes tests.

Maybe the most serious outcome of these negative effects on student motivation is that students may drop out of school altogether. Although students drop out for various reasons, high school graduation exams appear to increase the number of student retentions, which has increased the dropout rate (Amrein & Berliner, 2003; Haney, 2000; Jacob, 2001). Testing has increased retention rates by requiring students to pass tests to be promoted to the next grade and by pressuring some teachers to retain students who they doubt will pass the tests in the following year without being retained (Amrein & Berliner, 2003; McNeil, 2000). Retaining more students has likely increased dropout rates due to the fact that students who are retained are significantly more likely to drop out of school (Goldschmidt & Wang, 1999). Sadly, some teachers choose to spend less attention on students who are not likely to pass the tests, focusing instead on the "bubble kids" who can pass with a little extra help and who will give the teacher and school the biggest return on their investment (Booher-Jennings, 2006). Exactly who is dropping out and to what extent has been hotly debated because researchers have used different sets of data and different methods for calculating dropout rates (see Bracey, 2006, for a discussion). Consequently, it is difficult to state with certainty the extent of the dropout problem, although there is mounting research to suggest that testing policies have had an adverse effect on it (Wheelock, 2003).

Many *teachers* have also experienced increased stress from the pressure of the tests which has led some teachers to report negative attitudes

towards the profession, lower teacher morale, less enjoyment in their job, and an increase in teacher attrition (Center on Education Policy, 2006; DeBard & Kubow, 2002; Jones & Egley, 2004c). With respect to attrition, 85% of Texas teachers in one study agreed that some of the best teachers are leaving the field because of high-stakes testing (Hoffman, Assaf, & Paris, 2001) and 52% of teachers surveyed in two large Florida districts reported having thought about leaving the teaching profession in the past year (Tobin & Ave, 2006). Interestingly, some teachers who stay in the profession have reported that they wanted to transfer out of tested grades (Abrams, Pedulla, & Madaus, 2003). Moreover, some Florida administrators, especially those in rural schools, have reported that their school rating had negatively affected their ability to attract high quality teachers (Egley & Jones, 2004a). Teacher resignations and difficulty in teacher recruitment are two possible consequences that could severely affect the quality of education provided by schools. It is worth noting that the research about the effects of testing on teachers presented here is based on teacher perceptions, which is appropriate for assessing teachers' level of stress, attitudes towards the profession, morale, and enjoyment of their job. However, an analysis of actual attrition rates would be useful in verifying teachers' beliefs about teacher recruitment and retention.

Other Effects on Education

Educators, policy makers, and the general public have cited other negative effects of testing beyond those discussed previously in this paper. Among them is that testing is costly to implement and takes money away from more critical needs. For example, the cost for developing, administering, scoring, and reporting all components of the state testing program in Florida is about $42 million per year (Florida Department of Education, 2003). The Connecticut State Department of Education estimated that the costs of NCLB to the State Department of Education would be about $112.2 million in staff time and actual dollar outlay from 2002 through 2008 (Connecticut State Department of Education, 2005). Because the State of Connecticut only expected to receive $70.6 million from the federal government to cover these costs, the burden has fallen on the state to pay the $41.6 million difference. Consequently, the Commissioner of Education in Connecticut reported that that this money could have been spent on more critical education needs: "In sum, the $41.6 million in staff time and additional financial resources the State of Connecticut needs in order to meet specific NCLB demands

could be spent in much better ways–ways that would *truly* leave no child behind" (Connecticut State Department of Education, 2005, p. 29).

Finally, some educators believe that the testing has created a negative image of public education (Jones & Egley, 2004a, 2004c). As one teacher explained, "The [testing] makes schools look bad instead of celebrating many of their successes" (Jones & Egley, 2004c, *Theme 5*, para. 7). The following two newspaper headlines serve as examples of how the media can use test results to help create a negative image of public schools: "Few schools find reason to celebrate: Half of high schools get a D" (Tobin, 2004, June 16, p. B1; in Florida, schools are graded an A, B, C, D, or F); and "Pinellas schools losing their luster" (Tobin & Winchester, 2004, November 21).

UNINTENDED OUTCOMES FOR STUDENTS AT-RISK

One of the purposes of NCLB (2002) is to ensure that *all* children have a high-quality education. To this end, the policy requires that "assessment results and State progress objectives must be broken out by poverty, race, ethnicity, disability, and limited English proficiency to ensure that no group is left behind" (U.S. Department of Education, n.d., *Increased Accountability*). Clearly, the stated intent of NCLB is to focus on at-risk populations that might otherwise be forgotten or treated unfairly. A positive outcome of NCLB is that it has, in some cases, brought much needed attention to these disadvantaged groups. The purpose of the following sections is to examine some of the unintended consequences on these groups that have resulted from the implementation of high-stakes testing.

Students with Learning or Behavioral Problems

Much anecdotal evidence exists about the effects of high-stakes testing on students with disabilities, but the empirical evidence available is limited (see Ysseldyke et al., 2004 for a review). Many of the intended consequences of high-stakes testing appear to be positive in that some evidence suggests that for students with disabilities: scores have increased on high-stakes tests (Filbin, 2002; Gloeckler, 2001; Thompson & Thurlow, 2001), participation in the testing has increased (Schulte, Villwock, Whichard, & Stallings, 2001; Thompson & Thurlow, 2001), and educator and parental expectations have been raised (Gloeckler, 2001; Thompson & Thurlow, 2001).

Nonetheless, students with disabilities continue to underperform on high-stakes tests as compared to their nondisabled counterparts, regardless of the type of accommodations received (Horn, 2003; Koretz & Hamilton, 2001). One explanation for this underperformance is provided by Disability Rights Advocates (2001):

> One reason that high-stakes assessments have a discriminatory impact on students with learning disabilities is because often when the tests were developed, little or no attention was given to how the tests would impact learning disabled test takers. The sample population that is used by test developers to set the average scores for the tests usually does not include students with disabilities. When disabled students are included in the sample population, it is often unintentional, and the performance of these individuals is not separately tracked. Most testing publishers also do not give students with disabilities accommodations they need when testing a sample population, thus leading to a dearth of information and research about the true effect of an accommodation on a testing situation. (p. 3)

As a result of the lack of research about how testing accommodations affect test score validity, test publishers might label the accommodations as "non-standard" or "modifications" because they do not know how they affect test score validity (Disability Rights Advocates, 2001, p. 9). Even with accommodations, standardized testing conditions can unfairly disadvantage students with learning disabilities. For instance, multiple choice tests might not provide the sufficient context needed for dyslexic students who rely heavily on context to identify words (Disability Rights Advocates, 2001).

Many of the unintended outcomes noted previously in this article also apply to students with disabilities. For example, the effects of narrowing the curriculum to the subjects tested can have an equal if not greater negative impact on students with disabilities. As Ysseldyke et al. (2002) note: "The curriculum for students with disabilities may be narrowed in the sense that their remedial courses may prevent them from selecting other coursework that may help direct future vocational goals" (Nelson, 2002, p. 85).

Another often-cited unintended consequence for students with disabilities is an increase in anxiety, although many of these claims are anecdotal (Nelson, 2002, cited in Ysseldyke et al., 2002, p. 89). Even so, it is hard to ignore stories about students whose motivation and self-es-

teem have been negatively affected by high-stakes tests. Holbrook (2001) describes a few such cases of students who are learning disabled and are unable to pass the pass the Massachusetts Comprehensive Assessment System (MCAS) tests because of their lack of reading ability. Holbrook wrote that "For my fourth-graders, the present MCAS is a ridiculous waste of time, emotion, and self-esteem" (p. 784). Teachers such as Holbrook question why students with documented disabilities are required to take tests that they cannot pass given their disability or current level of cognitive development.

In sum, there appear to be some positive intended outcomes and some negative unintended outcomes associated with high-stakes testing for students with disabilities. Unfortunately, research has not addressed many other important questions related to high-stakes testing and students with disabilities, such as: How has testing affected the practice of "tracking" students with disabilities? Or, how has testing affected how students are provided supplementary services (e.g., summer school, remedial programs, mentors)?

Students from Economically Impoverished Families

Many reasons have been provided for why schools with students from economically impoverished families are less likely to demonstrate Adequate Yearly Progress (as defined by NCLB [2002]) than schools in more affluent areas (Smith, 2005). Of course, this issue is confounded by the fact that many low socioeconomic status (SES) students are also minorities, English Language Learners (ELLs), or both, all of whom tend to score lower on standardized tests. Nonetheless, factors associated with poverty create unique challenges for educators, as is evidenced by the fact that the gap between poor students and non-poor students has failed to close since the implementation of NCLB (Lee, 2006).

A clear correlation exists between poverty and low academic achievement (Berliner, 2006). So much so that English (2002) claims that the achievement gap will never be resolved because SES is a crucial variable in explaining test score variance. In this respect, high-stakes testing could be useful in highlighting the connection between poverty and low test scores and the need for more resources in these high-poverty schools. Unfortunately, what typically has happened is that high-stakes tests have been used to blame educators in high-poverty schools for their lack of success. As one Florida teacher explained, "The [high-stakes test] seems to be a way to make teachers scapegoats for

problems plaguing society" (Jones & Egley, 2004c, *Theme 5*, para. 7). Instead of blaming educators for poor test scores, Berliner (2006) argues that our whole society needs to be held as accountable for providing healthy children ready to learn. "One-way accountability, where we are always blaming the schools for the faults that we find, is neither just, nor likely to solve the problems we want to address" (Berliner, 2006, *Conclusion*, para. 3).

Few studies address the effects of testing specifically on the instruction of poor students. However, in studying school districts in New Jersey, Firestone et al. (2002) found that teaching to the test occurred most often in the poorest New Jersey districts. Similarly, in their study of Florida teachers, Jones and Egley (2004b) found that teachers at poorer schools reported spending more time teaching test-taking strategies. Furthermore, poor students are at a disadvantage because teacher quality has a large effect on student achievement (Darling-Hammond, 1999) and poor students are typically taught by less qualified teachers (e.g., Shen, Mansberger, & Yang, 2004).

Other negative effects of testing on low SES students relate to the fact that these students do not have access to resources in their home life to adequately prepare them for certain standardized tests. For instance, students from economically impoverished families may depend more heavily on their school to provide access to cultural, vocational, and enrichment activities because they have little access to these activities outside of school. Yet, as discussed previously, these types of activities have been disappearing from schools as educators narrow the curriculum to focus more on the basic skills in tested subjects such as reading, writing, and mathematics. Further, these limited experiences affect the background knowledge from which students can draw to answer test questions. The following example demonstrates the disadvantage to which low SES students can be subjected when their day-to-day life revolves around survival:

> Teachers gave several examples of the class bias in the SAT-9. Nicole mentioned that there might be a question about airplane travel, yet only two of her students have ever been to an airport. Mary mentioned a reading comprehension question on a passage about a woman who worked in the health profession. The students were required to make an inference on why the woman chose this job. The answer the test creator was looking for was that she enjoyed helping people. However, Mary noticed that her few students who could actually read and comprehend the passage chose

the answer, "because she needed the money." (Wright, 2002, *Social, cultural, and class bias*)

Schools attended by poorer students have also been impacted by high-stakes testing. Hodges (2002) noted that rural schools face unique challenges because they may lack educational service centers, have service centers that lack personnel, or are located in geographically isolated areas. This lack of services can lead to the following types of problems: (a) not having faculty trained in assessment, its administration, or its interpretation; (b) not having access to quality staff development to assess the causes of low test scores, to train teachers to alleviate these causes, or to follow up on success of new methods or programs; (c) not having time and personnel to align state and local standards to the test; and (d) not being able to prepare rural students for the tests or to provide them with information as to the importance of the tests (Hodges, 2002). Some school administrators in rural Florida also claimed that testing had a negative effect on their ability to attract high-quality teachers (Egley & Jones, 2004).

Students from Minority Groups

African-American and Hispanic students continue to score well below White students on academic achievement tests (National Center for Education Statistics, 2004). This finding has often been labeled the "achievement gap" because of the large gap between the higher test scores of White students and the lower test scores of some minority groups. Although some progress has been made in some areas (Center on Education Policy, 2006; Coley, 2003), the achievement gap has failed to substantially close over several decades. Discouragingly, NCLB has not had a significant effect on closing the racial gap (Lee, 2006) and the gap has actually widened for African-American students on the SAT (*The Journal of Blacks in Higher Education*, 2005).

In addition to factors associated with SES, African-American students might be at a disadvantage on high-stakes tests due to a phenomenon known as stereotype threat. Stereotype threat refers to the risk associated with confirming a negative stereotype based on group membership (Steele, 1997; Steele & Aronson, 1995). On high-stakes tests, African-American students might score lower than would be expected based on their abilities due to a fear that they will confirm the negative stereotype that African-American students score lower on high-stakes tests than White students (Kellow & Jones, 2005). For instance, when

ninth-grade African-American students were told that a test they were about to take was predictive of their ability on a high-stakes test, they scored lower than when they were told that African-American students typically scored the same on the test as White students (Kellow & Jones, 2005). These findings indicate that one factor in the achievement gap might be the stereotype threat felt by African-American students during high-stakes tests. Consequently, compared to White students, African-American students may be at a disadvantage because they are unable to demonstrate their true abilities on tests in which African-Americans have been shown to consistently score lower than White students. The negative outcome of stereotype threat is not only that African-American students score lower than would be expected on tests, but also that they may be retained or prevented from graduating high school based on these inaccurate test scores.

Students with Limited English Proficiency

Because English Language Learners (ELLs) are often minorities from low-income families (Cosentino de Cohen, Deterding, & Clewel, 2005), NCLB requires that these students are assessed for Adequate Yearly Progress (AYP) in three subgroups: the ELL subgroup, a racial/ethnic subgroup, and the low-income subgroup. Unfortunately, ELL students have consistently scored lower on high-stakes tests when compared to other students (Horn, 2003), which becomes a serious concern for educators who are worried about their school making AYP. As a result, states and districts consider the NCLB requirement for ELL students one of the law's greatest challenges because of the instructional time and resources that it consumes (Center on Education Policy, 2006). Much of the cost and difficulty in administering it is related to the fact that it requires one teacher or administrator to work with a single student (Center on Education Policy, 2006).

Others have questioned whether test scores for ELL students are reliable or valid (Adebi, 2003; Heubert & Hauser, 1999). Some educators are especially concerned that high-stakes tests do not accurately measure student learning and development (Jones & Egley, 2004c; Wright, 2002). In fact, Pedulla et al. (2003) found that nine out of ten teachers did not regard their state test as an accurate measure of what ELL students know or can do. A possible negative effect of such inaccurate test results is that it leads to frustrated students and teachers who do not believe that they have a fair chance to succeed in school according to these types of external measures.

Although many teachers have noted how testing has narrowed the curriculum and forced them to spend excessive time on test preparation, these concerns are paramount for some ELL teachers who have difficulty in designing individualized literacy instruction based on their professional judgment, experience, and expertise because of the amount of time it takes to implement district mandated reading programs (Wright, 2002). For instance, Wright (2002) describes how the curriculum in California shifted to a highly scripted one-size-fits-all model that relies on drills and worksheets leaving little time for reading. The curriculum was adopted because its use had raised student test scores on the SAT-9 in Texas and other parts of California. Such findings have led Wiley and Wright (2004) to conclude that "current federal education policy for language-minority students in need of English language development no longer mandates, nor even encourages, bilingual education" (p. 162).

CONCLUSION

In this article, I discussed several of the unintended outcomes of testing that have been reported, either empirically or anecdotally. It is difficult to generalize the findings from any one state, district, school, or classroom to another because contextual factors mediate the extent of these effects. Even within states, student and educator perceptions can vary significantly. Nonetheless, patterns in the data are beginning to emerge that can help us better understand the effects of high-stakes testing on public education. Unfortunately, many of the outcomes of testing have been negative. Certainly, positive effects of testing have been noted; however, we must continually ask whether the benefits outweigh the negative effects.

Because high-stakes testing has placed a tremendous focus on a single test score as a measure of a student's ability, one of my aims in writing this article was to present some of the many factors that can affect a student's test score. School psychologists should consider these factors when making decisions about students and when discussing the results of high-stakes tests with parents and other educators. School psychologists can use their position to educate students, parents, other educators, and the general public about the limitations of using high-stakes test scores as the sole evaluation of a student's ability. They can advocate for using these scores in combination with other indicators of a student's ability, such as grades, classroom behaviors, and scores on other types of tests and measurements. The use of other measures is espe-

cially important for at-risk students given the multitude of factors that can affect their standardized test scores and the often limited test score reliability and validity information available for these subgroups. In using the results of high-stakes tests appropriately and in educating others about the need for multiple measures of students' abilities, school psychologists can help to lessen the intense focus on and importance of high-stakes test scores.

REFERENCES

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice, 42*(1), 18-29.

Adebi, J. (2003) Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8*(3), 231-258.

American Educational Research Association. (2004, December). *High-stakes testing in PreK-12 education*. Retrieved March 30, 2006, from http://www.aera.net/policyandprograms/?id=378

Amrein, A. L., & Berliner, D. C. (2003). The effects of high-stakes testing on students' motivation and learning. *Educational Leadership, 60*(5), 32-38.

Barksdale-Ladd, M., & Thomas, K. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education, 51*(5), 384-397.

Berliner, D. C. (2006). Our impoverished view of educational reform. *Teachers College Record, 108*(6), 949-995.

Booher-Jennings, J. (2006). Rationing education in an era of accountability. *Phi Delta Kappan, 87*(10), 756-761.

Bracey, G. W. (2006). Dropping in on dropouts. *Phi Delta Kappan, 87*(10), 798-799.

Bush, J. (2003). *Lessons learned in the Sunshine State*. Hoover Digest, 3. Retrieved April 18, 2006, from http://www.hooverdigest.org/033/bush.html

Catalanello, R. (2004, February 1). Kicking "FCAT essay" habit. *St. Petersburg Times*. Retrieved February 6, 2004, from http://www.sptimes.com.

Center on Education Policy. (2006, March). *From the capital to the classroom: Year 4 of the No Child Left Behind Act*. Retrieved March 30, 2006, from http://www.cep-dc.org/nclb/Year4/Press/

Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives, 10*(2). Retrieved April 18, 2005 from http://epaa.asu.edu/apaa/v10n2.html

Coley, J. (2003). *Growth in school revisited: Achievement gains from the fourth to the eighth grade*. Princeton, NJ: Educational Testing Service.

Connecticut State Department of Education. (2005, March 2). *Cost of Implementing the Federal No Child Left Behind Act in Connecticut: State-Level Costs, Part I*. Retrieved July 19, 2006, from http://www.state.ct.us/sde/NCLB_Study_2_28_05.pdf

Cosentino de Cohen, C., Deterding, N., & Clewel, B. C. (2005). *Who's left behind: Immigrant children in high and low LEP schools.* Washington, DC: The Urban Institute

Darling-Hammond, L. (1999). *CTP research report: Teacher quality and student achievement: A review of state policy evidence*. Seattle, Washington: Center for the

Study of Teaching and Policy. Retrieved July 10, 2006 from http://depts.washington.edu/ctpmail/PDFs/LDH_1999.pdf

DeBard, R., & Kubow, P. K. (2002). From compliance to commitment: The need for constituent discourse in implementing testing policy. *Educational Policy, 16*(3), 387-405.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*, 105-115.

Disability Rights Advocates. (2001). *Do no harm–High stakes testing and students with learning disabilities*. Oakland, CA: LD Access Foundation. Retrieved May 3, 2006, from http://www.dralegal.org/publications/do_no_harm.php

Egley, R. J., & Jones, B. D. (2004). Rural elementary administrators' views of high-stakes testing. *The Rural Educator, 26*(1), 30-39.

English, F. W. (2002). On the intractability of the achievement gap in urban schools and the discursive practice of continuing racial discrimination. *Education and Urban Society, 34*(3), 298-311.

Filbin, J. (2002). No more field trips on test days: How students with disabilities benefit from the CSAP. *Inside Special Education, 3*(2), 2. Retrieved April 18, 2006, from http://www.cde.state.co.us/cdesped/download/pdf/inside_fall2001.pdf

Firestone, W. A., & Mayrowetz, D. (2000). Rethinking "high stakes": Lessons from the United States and England and Wales. *Teachers College Record, 102*(4), 724-749.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20*(2), 95-113.

Firestone, W. A., Monfils, L., Camilli, G., Schorr, R. Y., Hicks, J. E., & Mayrowetz, D. (2002). The ambiguity of test preparation: A multimethod analysis in one state. *Teachers College Record, 104*(7), 1485-1523.

Florida Department of Education. (2003). *FCAT Myths vs. Facts*. Retrieved July 13, 2006, from http://www.firn.edu/doe/sas/fcat/pdf/myths-facts.pdf

George, P. S. (2001). A + accountability in Florida? *Educational Leadership, 59*(1), 28-32.

Gilmer, K. R. (2002, June 13). 48 A's...and two F's. *St. Petersburg Times*, B1, B6.

Gloecker, L. C. (2001). The door to opportunity: Let's open it for everyone. *State Education Standard, 2*, 21-25.

Goldschmidt, P., & Wang, J. (1999). When can schools affect dropout behavior? A longitudinal multilevel analysis. *American Educational Research Journal, 36*(4), 715-738.

Goodlad, J. I. (1979). *What are schools for?* Bloomington, IN: Phi Delta Kappan.

Gordon, S. P., & Reese, M. (1997). High-stakes testing: Worth the price? *Journal of School Leadership, 7*, 345-368.

Haney, W. (2000, August 19). The myth of the Texas miracle. *Educational Policy Analysis Archives, 8*(41). Retrieved March 30, 2006, from http://epaa.asu.edu/epaa/v8n41/

Haney, W. (2002, May 6). Lake Woebeguaranteed: Misuse of test scores in Massachusetts, Part I. *Educational Policy Analysis Archives, 10*(24). Retrieved July 10, 2006, from http://epaa.asu.edu/epaa/v10n24/

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

HISD Connect (no date). *Teacher Performance Pay*. Retrieved March 29, 2006 from http://www.houstonisd.org/

Hodges, V. P. (2002). High stakes testing and its impact on rural schools. *Rural Educator, 24*(2), 3-7.

Hoffman, J., Assaf, L., & Paris, S. (2001). High-stakes testing in reading: Today in Texas, tomorrow? *The Reading Teacher, 54*, 482-492.

Holbrook, P. J. (2001). When bad things happen to good children: A special educator's views of MCAS. *Phi Delta Kappan, 82*(10), 781

Horn, C. (2003). High-stakes testing and students: Stopping or perpetuating a cycle of failure? *Theory Into Practice, 42*(1), 30-41.

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Education Evaluation and Policy Analysis, 23*(2), 99-121.

Jones, B. D., & Egley, R. J. (2004, April 15). *Learning to understand or learning to achieve? Effects of high-stakes testing on student learning.* Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Jones, B. D., & Egley, R. J. (2004a). Is testing the right direction? Administrators share their thoughts. *ERS Spectrum, 22*(3), 16-25.

Jones, B. D., & Egley, R. J. (2004b, April 15). *Learning to understand or learning to achieve? Effects of high-stakes testing on student learning*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Jones, B. D., & Egley, R. J. (2004c). Voices from the frontlines: Teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives, 12*(39). Retrieved August 23, 2004, from http://epaa.asu.edu/epaa/v12n39/

Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan, 81*(3), 199-203.

Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Kellaghan, T., Madaus, G. F., & Raczek, A. (1996). *The use of external examinations to improve student motivation*. Washington, D.C.: American Educational Research Association.

Kellow, J. T., & Jones, B. D. (2005). Stereotype threat in African-American high school students: An initial investigation. *Current Issues in Education, 8*(20). Retrieved October 25, 2005, from http://cie.asu.edu/volume8/number20

Koretz, D., & Hamilton, L. (2001). *The performance of students with disabilities on the New York Regents comprehensive examination of English* (CSE technical report 540). University of California, Los Angeles: Center for the Study of Evaluation. Retrieved March 25, 2006, from http://www.cse.ucla.edu/CRESST/pages/reports.htm

Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends*. Cambridge, MA: The Civil Rights Project at Harvard University.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic rewards: A test of the "Overjustification Hypothesis." *Journal of Personality and Social Psychology, 28*, 129-137.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge Kegan Paul.

National Association of School Psychologists (2003). *Position statement on using large scale assessment for high stakes decisions*. Retrieved July 13, 2006, from http://www.nasponline.org/information/pospaper_largescale.html

National Center for Educational Statistics (2004). *National assessment of educational progress: 2004 long-term trend assessment results.* Retrieved August 4, 2005, from http://nces.ed.gov/nationsreportcard

National Research Council. 2000. *How people learn*. Washington, D.C.: National Academy Press.

Nelson, J. R. (2002). *Closing or widening the gap of inequality: The intended and unintended consequences of Minnesota's Basic Standards Tests for Students with Disabilities*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

No Child Left Behind Act of 2001. (2002, January 8). Public Law 107-110, 115 Stat. 1425. Retrieved April 18, 2006, from http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston College: National Board on Educational Testing and Public Policy.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Upper Saddle River, NJ: Merrill Prentice Hall.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8-15.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders.* Boston: Allyn & Bacon.

Popham, W. J. (2004). *America's "failing" schools: How parents and teachers can cope with* No Child Left Behind. New York: Routledge Falmer.

Rapp, D. (2002). National board certified teachers in Ohio give state education policy, classroom climate, and high-stakes testing a grade of F. *Phi Delta Kappan, 84*(3), 215-218.

Reese, M., Gordon, S. P., & Price, L. R. (2004). Teachers' perceptions of high-stakes testing. *Journal of School Leadership, 14*(5), 464-496.

Reeves, E. B. (2000). High-stakes accountability and contextual effects: An empirical study of the fairness issue. *Research in the Schools, 7*(2), 49-58.

Schulte, A., Villwock, D. V., Whichard, S. M., & Stallings, C. (2001). High-stakes and expected progress standards for students with learning disabilities: A five-year study of one district. *School Psychology Review, 30*, 487-506.

Shen, J., Mansberger, N. B., & Yang, H. (2004). Teacher quality and students placed at risk: Results from the baccalaureate and beyond longitudinal study 1993-1997. *Educational Horizons, 82*(3), 226-235.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11.

Smith, E. (2005). Raising standards in American schools: The case of No Child Left Behind. *Journal of Education Policy, 20*(4), 507-524.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613-629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.

The Journal of Blacks in Higher Education (2005). The widening racial scoring gap on the SAT college admissions test. *The Journal of Blacks in Higher Education, 49*. Retrieved March 25, 2006, from http://www.jbhe.com/features/49_college_admissions-test.html

Thompson, S., & Thurlow, M. (2001). *2001 State special education outcomes: A report on state activities at the beginning of a new decade*. Minneapolis, MN: Univer-

sity of Minnesota, National Center on Educational Outcomes. Retrieved July 10, 2006, from http://education.umn.edu/NCEO/OnlinePubs/2001StateReport.html

Tobin, T. C. (2004, June 16). Few schools find reason to celebrate: Half of high schools get a D. *St. Petersburg Times*, pp. B1, B4.

Tobin, T. C. & Ave, M. (2006, May 14). Teachers troubled with job, poll says. *St. Petersburg Times*, pp. A1, A6.

Tobin, T. C., & Winchester, D. (2004, October 4). New FCAT push accelerates teaching pace. *St. Petersburg Times*, pp. A1, A6.

Tobin, T. C., & Winchester, D. (2004, November 21). Pinellas schools losing their luster. *St. Petersburg Times*, pp. A1, A10.

Triplett, C. F., Barksdale, M. A., & Leftwich, P. (2003). Children's perceptions of high stakes testing. *Journal of Research in Education, 13*(1), 15-21.

U.S. Department of Education (no date). *Executive Summary*. Retrieved April 18, 2006, from http://www.ed.gov/nclb/overview/intro/execsumm.html

Wheelock, A. (2003). Myopia in Massachusetts. *Educational Leadership, 61*(3), 50-54.

Wheelock, A., Bebell, D. J., & Haney, W. (2000). What can student drawings tell us about high-stakes testing in Massachusetts? *Teachers College Record*. Retrieved July 10, 2002 from http://www.tcrecord.org/Content.asp?ContentID = 10634

Wiley, T. G., & Wright, W. E. (2004). Against the undertow: Language-minority education policy and politics in the "Age of Accountability." *Educational Policy, 18*(1), 142-168.

Wilkins, J. (2000, April). *Characteristics of demographic opportunity structures and their relationship to school-level achievement: The case of Virginia's standards of learning*. Paper presented at the American Educational Research Association, New Orleans.

Wright, W. E. (2002). The effects of high-stakes testing on an inner-city elementary school: The curriculum, the teachers, and the English language learners. *Current Issues in Education, 5*(5). Retrieved March 25, 2006, from http://cie.asu.edu/volume5/number5

Yarbrough, T. L. (1999). *Teacher perceptions of the North Carolina ABC program and the relationship to classroom practice*. Ph.D. dissertation, University of North Carolina at Chapel Hill.

Ysseldyke, J., Nelson, R. J., Christenson, S., Johnson, D. R., Dennison, A., & Triezenberg, H. et al. (2004). What we know and need to know about the consequences of high-stakes testing for students with disabilities. *Exceptional Children, 71*(1), 75-94.

doi:10.1300/J370v23n02_05