High-Stakes Testing: Does It Increase Achievement?

Sharon L. Nichols

University of Texas at San Antonio

SUMMARY. I review the literature on the impact on student achievement of high-stakes testing. Its popularity as a mechanism for holding educators accountable has triggered studies to examine whether its promise to increase student learning has been fulfilled. The review concludes there is no consistent evidence to suggest high-stakes testing leads to increases in student learning. Some evidence suggests it may have a negative effect for some student groups and in some important subject areas (e.g., reading). Implications for future research and for the practice of school psychology are discussed. doi:10.1300/J370v23n02_04

Address correspondence to: Sharon L. Nichols, College of Education and Human Development Department of Counseling, Educational Psychology, and Adult Higher Education, University of Texas at San Antonio, 501 West Durango Blvd., Suite DB 4.342, San Antonio, TX 78207-4415 (E-mail: sharon.nichols@utsa.edu).

The author would like to thank Tom Good for providing extensive feedback on earlier versions of this paper. The author also thanks David Berliner,Gene Glass, Michael Karcher, and Jeremy Sullivan for their helpful feedback and editorial suggestions on an earlier draft. Lastly, the author thanks Michelle Lynde for her extensive editorial help in preparing this manuscript.

[Haworth co-indexing entry note]: "High-Stakes Testing: Does It Increase Achievement?" Nichols, Sharon L. Co-published simultaneously in *Journal of Applied School Psychology* (The Haworth Press, Inc.) Vol. 23, No. 2, 2007, pp. 47-64; and: *High Stakes Testing: New Challenges and Opportunities for School Psychology* (ed: Louis J. Kruger, and David Shriberg) The Haworth Press, Inc., 2007, pp. 47-64. 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_04

KEYWORDS. Accountability, academic achievement, high stakes tests, educational policy, achievement tests

INTRODUCTION

The passage of the No Child Left Behind (NCLB) Act in 2002 increased the practice of high-stakes testing in America's schools. Although high-stakes testing is not new (Amrein & Berliner, 2002a; Linn, 2000), never before has the practice been so widely applied. In supporting NCLB, politicians from both sides of the aisle enthusiastically endorsed high-stakes testing as the mechanism for holding administrators, teachers, and their students accountable for what they learn. But is it working? In the years leading up to NCLB and since its passage, several studies have examined the effects of high-stakes testing on student achievement.

In contrast to the literature on the mostly deleterious and *un*intended effects of high-stakes testing, which is substantial and largely indisputable (Amrein & Berliner, 2002a; Jones, Jones, & Hargrove, 2003; Neill, Guisbond, & Schaeffer, 2004; Nichols & Berliner, 2005; Orfield & Kornhaber, 2001; Ryan, 2004; Valenzuela, 2005), research on the relationship between high-stakes testing and its *intended* impact on achievement is sparse. Studies have varied widely in scope and design making it difficult to reach a single conclusion about the effects of high-stakes testing policy on student achievement. Rapid policy changes also have made it difficult to replicate earlier analyses. Nevertheless, in this review I offer some tentative conclusions regarding the efficacy of high-stakes testing policy for increasing student achievement.

The purpose of this review is to describe what is known to date about the impact of high-stakes testing policy on student achievement. I review a few of the more prominent studies in this area and discuss not only their findings, but also important methodological issues. This review culminates with work by myself and colleagues that describes a unique methodological approach to measuring high-stakes testing pressure to look at the effects of this pressure on student achievement (Nichols, Glass, & Berliner, 2006). Our work in combination with other studies lead to the conclusion that high-stakes testing has not been successful in increasing what students learn in school.

Rationale for High-Stakes Testing

The theory of action undergirding the practice of high-stakes testing is that when faced with large incentives and threats of punishment teachers will work harder and be more effective, students will be more motivated, and parents will become more involved (e.g., McDonnell, 2005; Raymond & Hanushek, 2003). More specifically it is commonly held that high-stakes testing will be effective because:

- teachers need to be held accountable through high-stakes tests to motivate them to teach better, particularly to push the laziest ones to work harder;
- students work harder and learn more when they have to take high-stakes tests;
- scoring well on the test will lead to feelings of success, while doing poorly on such tests will lead to increased effort to learn;
- high-stakes tests are good measures of an individual's performance, little affected by differences in students' motivation, emotions, language, and social status; and
- teachers will use test results to provide better instruction for individual students (Amrein & Berliner, 2002b, pp. 4-5).

In short, the pressure of doing well on a test, it is argued, will spur everyone into action, thus improving American public schools significantly (Haertel & Herman, 2005; Peterson & West, 2003; Phelps, 2005). Regardless of these common sense assumptions, the answer as to whether high-stakes testing works to improve student learning is less clear.

HIGH-STAKES TESTING AND STUDENT ACHIEVEMENT

A literature search on this topic yields a wide array of work that varies in scope, design, and emphasis (Herman & Haertel, 2005). I review some of the more notable studies that illustrate the main findings and important methodological concerns that arise when studying high-stakes policy implementation and impact.

Lake Wobegon

One early exchange examining the impact of high-stakes testing on student achievement occurred in the late 1980s when John Cannell re-

leased an acrimonious report that examined how districts and states had reported their Iowa Test of Basic Skills (ITBS) results. Known as the Lake Wobegon effect, Cannell's (1988) analysis pointed out that states had reported ITBS results where more than 50% of students were performing above average. Seeing this as a statistical improbability, he argued that it was the pressure of public reporting that compelled states and districts to manipulate the data to look "more favorable." Questions regarding Cannell's analytic approach prompted scholars to replicate his analysis. Many began by simply asking whether it was statistically possible that, as in Garrison Keillor's fictional community of Lake Wobegon, all of our students were performing above average?

Linn, Graue, and Sanders (1990) examined ITBS data carefully and found that Cannell was right: "The overall percent of students above the national median is greater than 50 in all of the elementary grades in both reading and mathematics" (Linn et al., 1990, p. 6). They also found that the average number of students above the national median at the elementary school level was higher in math across the three-year study period than in reading, going from a low of 58% in grade 4 (1985-1986) to a high of 71% in grade 2 (1987-1988). By contrast, in reading it ranged from 52% in grade 5 (1985-1986) to 60% in grade 3 (1987-1988). A similar, but less dramatic, difference was found at the high school levels.

Although there was relative consensus that inflated reporting had occurred, there was much less agreement on why. Had the pressures of public reporting influenced administrators and teachers to fabricate learning gains? Were students becoming more proficient? Another explanation, offered by Linn et al. (1990), proposed that inflated ITBS results could be explained as a statistical artifact that is the natural outcome of the process of re-norming. They argue that it is natural to expect students' test performance to rise when newer test scores are being compared to older norms in part because students and teachers become increasingly familiar with test items and objectives. And, analyses of ITBS trends demonstrate just that. Norm referenced test performance tends to rise as the time period between its current administration and when the norming data were collected increases. But, this upward trend dips sharply in the year when new, and often times more stringent, norms are collected. Linn et al.'s (1990) analysis casts a shadow of doubt on Cannell's (1988) position that educators were purposely manipulating their data to look more favorable.

But another explanation emerged. Under conditions of pressure (i.e., being evaluated publicly by student test results), teachers and principals

changed their behavior to focus instruction more intently on the test. Shepard (1990) collected interview and survey data from state officials regarding test and curriculum based instruction and found substantial reason to believe that increased pressure on the test compelled educators to engage in practices that could be considered as teaching to the test.

Although Linn et al.'s (1990) and Shepard's (1990) evidence was relatively strong, another analytic approach offered evidence that seemed to support more conclusively the contention that teaching to the test was occurring. Linn et al. (1990) compared ITBS scores with scores on the NAEP with the rationale that parallel increases in NAEP performance would be evidence of students' transfer of learning–evidence that student achievement gains were real. However, after statistically accounting for test related differences (issues of sampling, content), they found that ITBS scores rose higher and faster than scores on the NAEP. They argued that teaching-to-the test may have played a significant role in the test inflation worries of Lake Wobegon.

Although there remains no consensus on the Lake Wobegon studies, the empirical exchange unveiled an important issue that applies as well to current analyses of high-stakes testing and student achievement. Is it appropriate to use high-stakes test scores as evidence that high-stakes testing is working? Shepard (1990) persuasively showed that we must worry that as the stakes of testing rise, educators will focus more intently on preparing their students for it. This type of over preparation could compromise the validity of the test result. The importance of looking at a comparable no stakes test (such as the NAEP) for evidence that high-stakes policies are working to increase student learning was an important outcome of Lake Wobegon studies.

Texas Myth

During the 1990s, the accountability movement gained momentum and Texas was at the forefront, steadily increasing the stakes attached to testing of its students and teachers. Initially, Texas had received high praise and national accolades for the "success" of their policies as evidenced by increasing student achievement (Grissmer & Flanagan, 1998; Palmaffy, 1998). It appeared that accountability was working. However, the bubble burst when others examined their data and found that achievement increases were largely a "myth" (Haney, 2000) and not significantly different from achievement gains made in other states (Camilli, 2000). By looking carefully at reporting trends and again by

comparing achievement on Texas's state test (then the Texas Assessment of Academic Skills or TAAS) with achievement on the NAEP, Haney (2000) and others (e.g., Klein, Hamilton, McCaffrey, & Stecher, 2000) found that the alleged and highly publicized success of high-stakes tests for increasing student learning was erroneous, largely the result of the same problems that were identified during the Lake Wobegon exchange. Namely, increases in the TAAS were more likely to be the result of teaching to the test and other problems (dropping lower scorers from taking the test, miscounting number of dropouts). Evidence seems to be mounting from Texas and elsewhere (e.g., Kentucky, see Koretz & Barron, 1998; Massachusetts, see Haney, 2002) that test validity is seriously compromised when high-stakes decisions are attached to test score performance (Haney, 2000; Pedulla et al., 2003).

Chicago's End to Social Promotion

During the 1996-1997 school year, in their quest to end social promotion, the Chicago Public School (CPS) district began tying serious consequences to ITBS.[1] This marked the first year students in grades 3, 6, and 8 could be held back for inadequate performance on a test and teachers could be reassigned or dismissed for their students' inadequate performance. Jacob (2002) and colleagues (Roderick, Jacob & Byrk, 2002) examined whether this policy had the intended effect of increasing student achievement by analyzing ITBS scores before and after the implementation of the policy. Jacob (2002) found significant increases on ITBS following the implementation of high-stakes testing. It appeared as if achievement levels rose and that perhaps the policy had something to do with it. But did it? Follow-up item level analyses revealed that ITBS math gains were largely the result of improvements on computation and number concept skills and not higher-level thinking skills such as problem solving and data interpretation. This finding raises questions about the validity of the argument that high-stakes testing increases learning. Here it appears as if it worked only to increase basic skills that are susceptible to teaching to the test practices.

Roderick, Jacob and Byrk (2002) also examined ITBS performance following the introduction of high-stakes testing in Chicago, looking specifically at achievement following the "gateway" school years when students could be held back for low test performance at grades 3, 6, and 8. They found that the introduction of high-stakes testing in Chicago had a varying effect on achievement. For example, they found that in

reading, the lowest achieving students benefited by the policy (achievement went up). However, the opposite was found in math where high achieving students benefited most. They also found evidence of school-level effects where students in low performing schools showed larger gains in achievement after policy implementation than students of similar skill levels in better performing schools. Thus, it appears that in CPS, high-stakes testing had a mixed effect. Sometimes it was associated with gains made by low achievers, and other times it is was associated with gains made by high achievers. Similarly, sometimes it was related to gains among third graders and other times among sixth graders. Of course, follow-up studies are needed to see if these findings are sustained; however, from this study alone, the varied pattern of results makes it difficult to draw conclusions about how high-stakes testing impacts student learning and teacher instruction.

High-School Graduation Exams

Some have investigated the impact of high school graduation exams (tests students must pass in order to receive a diploma) on student achievement. Jacob (2001), for example, looked at twelfth-grade achievement in reading and math as reported on the National Educational Longitudinal Survey (NELS) in states with and without high school graduation exams. After accounting for prior achievement and other background characteristics (e.g., SES, ethnicity), Jacob found that the institution of high school graduation exams was not related to student achievement. The only exception was with lower achieving students in reading where there emerged a slightly positive effect associated with high school graduation exams. But, Jacob (2001) also found that states with high school graduation exams had more dropouts than those without such exams.

Marchant and Paulson (2005) looked at the effect of high school graduation exams on state level graduation rates, aggregated SAT scores, and individual student SAT scores. By comparing graduation rates and SAT scores in states with a graduation exam against states without a graduation exam, they found that states with graduation exams had lower graduation rates and lower aggregate SAT scores. They also found that the requirement of a high school graduation exam had a negative relationship with individual student SAT performance. Importantly, conclusions from this study are made with caution due to study limitations (e.g., reliability of graduation rate calculation, selectivity of SAT takers).

RESEARCH EVOLVES WITH CHANGING POLICY

Amrein and Berliner (2002b) triggered the most recent debate regarding the impact of high-stakes testing policy. They looked at state-level NAEP achievement trends over time since 1990. Using time trend analysis, Amrein and Berliner (2002b) looked at NAEP trajectories before and after high-stakes testing policies were introduced and compared them with a national average. This approach allowed them to determine if significant increases in fourth and eighth-grade NAEP performance occurred as a result of high-stakes testing policies being implemented. Across each of the 28 states included in their study, they found a random pattern of effects. Sometimes math performance went up; sometimes it went down. Similar results were found for reading performance. Sometimes gains were found in fourth grade, sometimes in eighth.

Rosenshine (2003) reanalyzed the same data set utilized by Amrein and Berliner (2002b) using a different design[2] and found that the overall average NAEP scores of states with high-stakes testing rose more rapidly than the average scores in states without any programs. But, when he looked at trends at the individual state level, there was no consistent effect detected. Rosenshine (2003, p. 4) concluded that "although attaching accountability to statewide tests worked well in some high-stakes states it was not an effective policy in all states." In a follow-up response to Rosenshine (2003), Amrein-Beardsley and Berliner (2003) used his design approach, but also included in their analysis NAEP exclusion rates.[3] They concluded that although states with high-stakes tests seemed to outperform those without high-stakes tests on the fourth-grade math NAEP exams, this difference disappeared when they controlled for NAEP exclusion rates.

Braun (2004) also critiqued Amrein and Berliner (2002b) on methodological grounds. In his analysis of fourth- and eighth-grade math achievement (he did not look at reading) from 1992 to 2000, he found that when standard error estimates are included in the analyses, NAEP gains were greater in states with high-stakes testing for eighth-grade math than in those without in spite of exclusion rate differences. However, the impact of high stakes testing is much lower when it comes to fourth grade math achievement and it is almost absent when looking at cohort achievement trends (1992 fourth-grade math and 1996 eighth-grade math; 1996 fourth-grade math and 2000 eighth-grade math). Cohort analyses are important because they minimize the validity threats due to selection bias in the groups compared, time, or experience. Stu-

dents tested in fourth grade and then in eighth grade four years later theoretically have experienced the same kinds of changes in instruction and the same increases in high-stakes testing pressure. Thus, assuming it may take time for testing pressures to take effect, a significant finding of high-stakes testing producing greater achievement among a cohort of students would be a robust confirmation of the policy's impact.

Measuring High-Stakes Testing Policy on a Continuum

Carnoy and Loeb (2002) were among the first to craft an index that rated states along a continuum of accountability "strength" designed to measure the level of pressure of high-stakes testing (see Appendix A, Carnoy & Loeb, 2002). However, their innovative index (scale of 0-5) had some measurement problems in that the numerical distinctions were relatively vague. For example, to receive the highest strength of accountability score they note, "States receiving a 5 had to have students tested in several different grades, schools sanctioned or rewarded based on student test scores, and a high school minimum competency test required for graduation. Other states had some of these elements, but not others" (Carnoy & Loeb, 2002, p. 14).

Carnoy and Loeb (2002) examined how their accountability index related to NAEP achievement. Their findings were mixed. They found a significant increase in eighth grade math performance (among White, Black and Hispanic students) as a result of increased accountability pressure. By contrast, the increases for fourth-grade math performance were much smaller for Black and Hispanic students and nonexistent for White students. Thus, like Braun (2004), Carnoy and Loeb (2002) found that eighth grade math is positively related to increases in high-stakes testing pressure. Importantly, their analysis focused only on 1996-2000 math performance and did not look at progress before or after that time period, or in any other subject area.

Hanushek and Raymond (2005) used a regression model to estimate accountability as a function of how long states had enacted high-stakes testing policy. Their analysis looked at fourth through eighth grade NAEP changes across Black, Hispanic and White student groups. They found that the introduction of state accountability had a positive impact on student performance overall. But when disaggregated by ethnicity, they found that NAEP increases were much lower for Black and Hispanic students than for White students. Hanushek and Raymond conducted other analyses with similar results. They concluded that consequential based policy has a positive impact on NAEP achievement for some groups but

not others. However, they caution the reader that theirs is a "blunt" measure of accountability that does not take into consideration state level variation in how accountability policy is enacted.

Our Measure of High-Stakes Testing Pressure

My colleagues and I were dissatisfied with extant research primarily because of weaknesses in their approaches to the measurement of high-stakes testing policy, largely explained by a rapidly changing political climate. States have been quickly adopting and transforming their high-stakes testing policies making it difficult to measure and isolate their effects. At the time of our work, every state had adopted some form of high-stakes testing. Because every state then had some form of high-stakes testing policy, it was impossible to make group comparisons between states with and without such policies. Although Carnoy and Loeb (2002) and Hanushek and Raymond (2005) created scales for capturing state level accountability differences, we felt these measures were inadequate for accounting for differences in policy implementation and impact on schools, students, and teachers.

We created an empirically derived rating scale that captured a more differentiated version of testing pressure embedded in 25 states' accountability systems (only states with complete or almost complete NAEP participation since 1990 were included). The determination of our Assessment Pressure Rating (APR) relied on a set of portfolios constructed to describe in as much detail as possible the past and current character of accountability practices for each state. These portfolios included a wide range of documents describing the politics, legislative activity, and impact of a state's high-stakes testing program as well as newspaper articles that served as a proxy for legislative implementation and impact (a full description of the selection strategy and portfolio examples can be found elsewhere, Nichols, Glass, & Berliner, 2006). Using these portfolios, we enlisted the help of over 300 graduate level student participants, each of whom rated a single pair of states. The method of "comparative judgments" was adopted for scaling our study states along a hypothetical continuum of high-stakes testing pressure (Torgerson, 1960).[4]

Our Study Findings

Our analytic approach included a series of correlations looking at APR and fourth and eighth grade NAEP in math and reading (overall

and disaggregated by student ethnicity). We also did a series of antecedent-consequence analyses where we correlated earlier changes in high-stakes testing pressure (e.g., from 1994 to 1998) with later achievement changes (e.g., from 1998 to 2002), thus providing a relatively robust estimation of a causal relationship (see Gujarati, 1995). We found that pressure was related only to fourth grade math achievement; the greater the pressure, the higher the math performance for all groups of students. Second, among the dozens of correlations looking at eighth grade math, results were inconsistent (some were positive, some negative, most were absent). Third, there was little impact on reading achievement at either the fourth or eighth grade levels. However, in the few instances the relationship was statistically significant, the outcome was mostly negative suggesting high-stakes testing pressure may be eroding reading performance especially in the fourth grade. Thus, the pressure to improve teaching and learning through applying sanctions based on test results produced test score gains only where drill on basic skills might raise achievement, namely elementary school arithmetic. Lastly, like Braun (2004), we found no link between high-stakes testing pressure and cohort achievement in math or reading.

CONCLUSION

Overall, the findings from the most rigorous studies on high-stakes testing do not provide convincing evidence that high-stakes testing has the intended effect of increasing student learning. Moreover, the modest gains found in some studies should be viewed with caution since the findings indicate that increases in achievement could be the result of teaching to the test. Of course, as others have noted (e.g., Crocker, 2005; Mathews, 2006), teaching to the test in some form is desirable. In preparing students for a test, it seems reasonable that instruction will be aligned with the objectives that will be covered on the test. But teaching becomes counterproductive when academic activities are geared specifically for students to do better on a test. This is especially true when it comes at the cost of other kinds of instruction or subject matter coverage. Studies that consider performance on NAEP suggest that by and large, high-stakes testing does not lead to "real" learning gains, but rather manufactured ones that are more likely the result of greater attention to the material that will be tested.

The Amrein and Berliner (2002b), Rosenshine (2003), and Braun (2004) exchange highlights a few important issues. First, findings were

simply too inconsistent to conclude high-stakes testing has any type of systematic effect (positive or negative) on learning. A second issue is the importance of including exclusion rates in analyses. Any time averaged test results are reported, one must ask the question of who participated in the testing and who did not. Evidence increasingly suggests that as pressure to perform increases, the lower test scorers are more likely to be removed from taking the test therefore inflating average test scores (Clarke et al., 2003; Nichols & Berliner, 2005). Any analysis attempting to connect high-stakes testing with achievement must account for the test taking pool. A third issue is how high-stakes testing is measured. Rosenshine (2003) found that achievement, when averaged across states, was higher in high-stakes testing states than in non stakes states. However, these findings disappeared when the data were disaggregated by state suggesting that implementation differences probably matter. That is, states with the same amount pressure may yield increases or decreases in learning–a result that could be attributed to the way the policy is implemented and received.

Implications and Future Directions

School psychologists must be aware of the inherent limitations of studies investigating high-stakes testing impact on student achievement. From Lake Wobegon, we learned about the potential problems associated with using the high-stakes test itself as a measure of high-stakes testing policy effectiveness. As Shepard (1990) and many others have now illustrated (e.g., Nichols & Berliner, in press-a; Pedulla et al., 2003), we must be suspicious of the validity of test scores when that test is used for making decisions about teachers and students. Studies of high-stakes testing policy must turn to "audit" tests such as the NAEP for an indication that learning has occurred.

A second critical element of these studies is how high-stakes testing policy is measured. Prior to NCLB and when only a selection of states had implemented accountability-based testing, it was appropriate to employ two group designs (comparing achievement trends in states with high-stakes against those without). However, now that all states employ some type of high-stakes testing policy, this approach is no longer relevant. Instead, researchers must find ways to measure the level of differentiation in their high-stakes testing policy from state to state as did Carnoy and Loeb (2002), Hanushek and Raymond (2005), and Nichols, Glass, and Berliner (2006). From these studies, we see that not all high-stakes testing policies are created equal. Some states' account-

ability practices are harder or more intense than others. The difference, however, is how this notion of state-level pressure was conceptualized. Carnoy and Loeb looked at the number of laws each state had, Hanushek and Raymond (2005) considered the length of time laws were on the books, and Nichols, Glass, and Berliner (2006) attempted to account for the implementation and impact of these laws. Thus, school psychologists must be critical consumers of the research on high-stakes testing and carefully weigh the efficacy of the conceptualization and measurement of high-stakes testing policy.

Although some of the gains on state developed standardized tests may be due to increased learning, there are data to suggest that much of the gains are a result of drilling and test preparation (Darling-Hammond & Rustique-Forrester, 2005; Jones & Egley, 2004; Shepard, 1990; Taylor et al., 2003). Thus, school psychologists could play a significant role in helping teachers avoid the pitfalls of teaching to the test when they are under pressure. The Center on Education Policy recommends that teachers avoid: (a) using the actual test questions from current test form and teaching students the answers, (b) giving students actual test questions for drill, review, or homework, or (c) copying distributing, or keeping past versions of a test that have not been officially released as a practice exam (Kober, 2002). As "obvious" as these tips seem, I remind readers that as the pressures of doing well on tests increase, so will the temptation to teachers and their students to engage in practices that might encroach on these warnings (Nichols & Berliner, in press-b). Thus, school psychologists must be prepared to support teachers when the stress increases and these temptations become more salient.

Additionally, school psychologists could also be helpful in assisting teachers to cope with their own test related anxiety and that of their students (see Kruger, Wandle, & Struzziero, 2007). As the pressure on teachers increases, they will find themselves increasingly in positions where the morality of their decision making will be challenged. For example, as NCLB marches on, more and more students are likely to fail the high-stakes test, putting more and more of our students in a vulnerable position where they may be retained in school, denied a diploma or scholarship monies. Teachers will be challenged by these dilemmas that ask them whether they should cross the line in order to "save" a struggling student or in order to save their job. These situations will inevitably trigger many difficult questions. Is following testing protocols worth it when students are so stressed out? How hard should I work to keep a struggling student in school when their lowered test score may

threaten my livelihood? These are tough questions all teachers and school psychologists will have to face.

Additionally, school psychologists must be armed and ready to face a growing number of students with test related anxiety. As the stakes rise, so too does the likelihood for failure for more of our students. Thus, school psychologists must be prepared to help students cope with and manage their anxiety and fears related to taking any high-stakes test. As part of this, it seems important that school psychologists help all students and all teachers realize that the goals of schooling are much bigger than what the performance on a single test suggests. As all educators are fully aware, students are social beings with complex lives (McCaslin & Good, 1996). In helping students and teachers cope with test related anxieties, it is important to help them gain perspective on the overall importance of the test in the long run. I suspect this will become increasingly harder if we continue to increase the stakes to teachers and their students. Still, school psychologists must be prepared to counsel students that their livelihoods do not rest solely on the performance on a single test.

Although more research is needed to unpack the relationship between high-stakes testing and student achievement, the evidence available provides ample reason to suspect it is not having the intended effect of increasing what students learn. By contrast, the literature is replete with stories about how the pressures of tests transform instruction and curricula to focus almost entirely on preparing for the test (Nichols & Berliner, in press-a). If we continue to hold students and their teachers accountable for performance on a single test, we run the risk of narrowing students' schooling experiences and thereby transforming public education into nothing more than a drill and kill set of exercises and demands.

School psychologists are in a position to play a significant role in voicing the problems and pitfalls of high-stakes testing to administrators, teachers, and parents. Through their voice, perhaps more will come to understand the significant limitations of test scores for representing what students know. Perhaps even more importantly, however, is the need for school psychologists to join with others to voice their concerns to those who set policy. The more often we share with our representatives personal stories of the effects of high-stakes testing on schools, teachers, and students, the more impact we may have for improving the way our students are assessed and our teachers are evaluated.

NOTES

1. NAEP data are not available at the district level.

2. Rosenshine wanted to address what were viewed as flaws in Amrein and Berliner's analysis. Instead of comparing states with high-stakes testing policies against a national average, Rosenshine compared states with high stakes testing policy with those that had no such policies.

3. Exclusion rates are defined as those students excluded from the assessment because "school officials believed that either they could not participate meaningfully in the assessment or that they could not participate without assessment accommodations that the program did not, at the time, make available. These students fall into the general categories of students with disabilities (SD) and limited-English proficient students (LEP). Some identified fall within both of these categories." From Pitoniak, M. J., & Mead, N. A. (2003, June). Statistical methods to account for excluded students in NAEP. Educational Testing Service, Princeton, NJ. Prepared for U.S. Department of Education; Institute of Education Sciences, and National Center for Education Statistics; p. 1. Retrieved February 14, 2005 from http://nces.ed.gov/nationsreportcard/pdf/main2002/statmeth.pdf

4. APR Results (lower number represents lower pressure): KY = .54, WY = 1.00, CT = 1.60; HI = 1.76, ME = 1.78, RI = 1.90, MO = 2.14, CA = 2.56, AK = 2.60, UT = 2.80, MD = 2.82, AL = 3.06, VA = 3.08, WV = 3.08, MA = 3.18, SC = 3.20, NM = 3.28, AZ = 3.36, GA = 3.44, TN = 3.50, LA = 3.72, MS = 3.82, NY = 4.08, NC = 4.14, TX = 4.78.

REFERENCES

Amrein, A.L., & Berliner, D.C. (2002a). High-Stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved January 7, 2004, from http://epaa.asu.edu/epaa/v10n18/

Amrein, A.L., & Berliner, D.C. (2002b). *The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP test results in states with high school graduation exams*. Retrieved January 7, 2004, from Arizona State University, Education Policy Studies Laboratory Web site: http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-126- EPRU.pdf

Amrein-Beardsley, A., & Berliner, D. (2003, August 4). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. *Education Policy Analysis Archives, 11*(25). Retrieved February 5, 2005, from http://epaa.asu.edu/epaa/v11n25/

Braun, H. (2004, January 4). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives, 12*(1). Retrieved February 5, 2005, from http://epaa.asu.edu/epaa/v12n1/

Camilli, G. (2000). Texas gains on NAEP: Points of light? *Educational Policy Analysis Archives, 8* (42). Retrieved September 21, 2006, from http://epaa.asu.edu/epaa/v8n42.html

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice, 7*(2), 5-9.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from interviews with educators in low-, medium-, and high-stakes states*. Retrieved January 7, 2004, from Boston College, National Board on Educational Testing and Public Policy Web site: http://www.bc.edu/research/nbetpp/statements/nbr1.pdf

Crocker, L. (2005). Teaching for the test: How and why test preparation is appropriate. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 159-174). Mahwah, NJ: Erlbaum.

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement. The 104th Yearbook of the National Society for the Study of Education, Part II* (pp. 289-319). Malden, MA: Blackwell.

Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas*. Washington, D.C.: National Education Goals Panel.

Gujarati, D. N. (1995)(3rd ed.), *Basic econometrics*. New York: McGraw Hill.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement. The 104th Yearbook of the National Society for the Study of Education, Part II* (pp. 1-34). Malden, MA: Blackwell.

Haney, W. (2000, August). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved February 5, 2005, from http://epaa.asu.edu/epaa/v8n41/

Haney, W. (2002). Lake Woebeguaranteed: Misuse of test scores in Massachusetts. *Educational Policy Analysis Archives, 10*(24). Retrieved September 22, 2006, from, http://epaa.asu.edu/epaa/v10n24/

Hanushek, E., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Herman, J. L., & Haertel, E. H. (Eds.). (2005). *Uses and misuses of data for educational accountability and improvement. The 104th Yearbook of the National Society for the Study of Education, Part II*. Malden, MA: Blackwell.

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

Jacob, B. (2002, May). *Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools*. (National Bureau of Economic Research Working Paper No. W8968). Abstract retrieved June 3, 2006, from http://ssrn.com/abstract = 314639

Jones, B. D., & Egley, R. J. (2004, August 9). Voices from the frontlines: Teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives, 12*(39). Retrieved December 2, 2004, from *http://epaa.asu.edu/epaa/v12n39/*

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

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000, October 26). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8*(49). Retrieved February 22, 2005, from http://epaa.asu.edu/epaa/v8n49/

Kober, N. (2002, June). Teaching to the test: The good, the bad, and who's responsible. *Test talk for leaders, 1*. Retrieved June 30, 2006, from http://www.cep-dc.org/testing/testtalkjune2002.pdf

Koretz, D., & Barron, S. (1998). The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS). Rand Corporation.

Kruger, L. J., Wandle, C. & Struzzieto, J. (2007). Coping with the Stress of High Stakes Testing. *Journal of Applied School Psychology, 23* (2), 109-128.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5-14.

Marchant, G. J., & Paulson, S. E. (2005, January 21). The relationship of high school graduation exams to graduation rates and SAT scores. *Education Policy Analysis Archives, 13*(6). Retrieved June 30, 2006 from http://epaa.asu.edu/epaa/v13n6/.

Mathews, J. (2006, February 20). Let's teach to the test. *The Washington Post*, p. A21.

McCaslin, M., & Good, T. (1996). The informal curriculum. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 622-673). New York: Macmillan.

McDonnell, L. M. (2005). Assessment and accountability from the policymaker's perspective. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement. The 104th Yearbook of the National Society for the Study of Education, Part II* (pp. 35-54). Malden, MA: Blackwell.

Neill, M., Guisbond, L., & Schaeffer, B. (with Madison, J. & Legeros, L.). (2004). *Failing our children: How "No Child Left Behind" undermines quality and equity in education and an accountability model that supports school improvement*. Cambridge, MA: Fairtest.

Nichols, S., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Nichols, S., & Berliner, D. C. (2007). The pressure to cheat in a high-stakes testing environment. In E. Anderman and T. Murdock (Eds.), *Psychological Perspectives on Academic Cheating* (pp. 289-312). NY: Elsevier.

Nichols, S. & Berliner, D. C. (2005, March). *The inevitable corruption of indicators and educators through high-stakes testing*. Retrieved March 23, 2005, from the Great Lakes Center Web site: http://www.greatlakescenter.org/pdf/EPSL-0503-101-EPRU.pdf

Nichols, S., Glass, G. V., & Berliner, D.C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1). Retrieved June 9, 2006, from http://epaa.asu.edu/epaa/v14n1/

Orfield, G. & Kornhaber, M.L. (Eds.). (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

Palmaffy, T. (1998). The Gold Star State: How Texas jumped to the head of the class in elementary school achievement. *Policy Review, 88*, p. 30-38.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003, March). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Retrieved January 7, 2004, from Boston College, National Board on Educational Testing and Public Policy Web site: http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

Peterson, P. E., & West, M. R. (Eds). (2003). *No Child Left Behind? The politics and practice of school accountability*. Washington, DC: Brookings Institute.

Phelps, R. P. (Ed.) (2005). *Defending standardized testing*. Mahwah, NJ: Erlbaum.

Phillips, G. W. (1990). The Lake Wobegon effect. *Educational Measurement: Issues and Practice, 9*(3), 3, 14.

Raymond, M. E., & Hanushek, E. A. (2003). High-stakes research [Electronic version]. *Education Next 3*(3), pp. 48-55.

Roderick, M., Jacob, B. A., & Byrk, A. S. (2002). The impact of high-stakes testing in Chicago on student achievement in promotional gate grades. *Educational Evaluation and Policy Analysis, 24*(4), 333-357.

Rosenshine, B. (2003, August 4). High-stakes testing: Another analysis. *Education Policy Analysis Archives, 11*(24). Retrieved January 7, 2004, from http://epaa.asu.edu/epaa/v11n24/

Ryan, J. E. (2004). The perverse incentives of the No Child Left Behind Act. *New York University Law Review, 79*, 932-989.

Shepard, L. A. (1990). Inflated test scores gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9*(3), 15-22.

Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2003). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost*. (CSE Technical Report 588: CRESST/CREDE/University of Colorado at Boulder). Los Angeles: University of California.

Torgerson, W. S. (1960). *Theory and methods of scaling*. New York: John Wiley.

Valenzuela, A. (Ed.). (2005) *Leaving children behind: How "Texas-style" accountability fails Latino youth*. Albany, NY: State University of New York Press.

doi:10.1300/J370v23n02_04