Using Data from High-Stakes Testing in Program Planning and Evaluation

Jeffery P. Braden

North Carolina State University

SUMMARY. This article intends to help school psychologists understand the nature of high stakes tests, methods for analyzing and reporting high stakes test data, standards for tests and program evaluation, and application of appropriate practices to program planning and evaluation. Although it is readily acknowledged that high stakes test data are not sufficient for effective program planning and evaluation, the availability of test results, and their salience for federally mandated accountability programs, argues in favor of using such data for program planning and evaluation. A decision-making model, which begins with high stakes test data, but also requires additional data from teachers and classrooms, is proposed to help practitioners evaluate program effectiveness, and make plans to improve student outcomes. doi:10.1300/J370v23n02_08

Address correspondence to: Jeff Braden, Department of Psychology, PO Box 7650, North Carolina State University, Raleigh, NC 36795-7650 (E-mail: jeff_braden@ncsu.edu)

[Haworth co-indexing entry note]: "Using Data from High-Stakes Testing in Program Planning and Evaluation." Braden, Jeffery P. Co-published simultaneously in *Journal of Applied School Psychology* (The Haworth Press, Inc.) Vol. 23, No. 2, 2007, pp. 129-150; and: *High Stakes Testing: New Challenges and Opportunities for School Psychology* (ed: Louis J. Kruger, and David Shriberg) The Haworth Press, Inc., 2007, pp. 129-150. 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_08

KEYWORDS. High stakes tests, program planning, evaluation, school psychology, data-based decision making

The phrase "high stakes test" is used often in public and professional literature to describe annual testing programs states conduct to comply with No Child Left Behind Act of 2001 (NCLB). Classic definitions of "stakes" or test consequences emphasize the immediacy and social salience of test consequences as factors in determining what makes stakes high or low (Heubert & Hauser, 1999). The test itself does not have any stakes per se; rather, it is the consequences institutions assign to test outcomes that determine the stakes of the test (Braden, 2002; Braden & Tayrose, in press).

In this light, annual testing required by NCLB is a low stakes enterprise for students, and a medium- to high-stakes enterprise for teachers and schools. NCLB does not require any consequences for test takers beyond the requirement that states must inform parents of their children's test scores. Currently, 23 of the states attach graduation decisions to high school tests, and only 8 use tests for grade promotion purposes (Education Week, 2006). Consequences for educators and schools are more profound, although consequences for failing to meet adequate yearly progress (AYP) goals for multiple consecutive years only apply to schools receiving Title I funds under NCLB. However, some states attached additional consequences to AYP determinations; 37 states provide additional assistance to any school identified in need of improvement, and 16 states provide rewards or incentives for schools that meet performance or improvement targets. Only 5 states withhold funds from schools failing to make AYP, suggesting that the assumption that NCLB tests lead to less funding for the lowest performing schools is inaccurate. In fact, all 5 states that withhold funds initially provide additional funds to improve failing schools (Education Week, 2006).

It is ironic to point out that, in an issue devoted to high stakes tests, tests themselves do not have stakes. Rather, it is the consequences that social institutions attach to test results that create "stakes," and stakes vary by stakeholders. The tests states use to implement NCLB mandates have low stakes for most students (i.e., less than half of states mandate consequences for students on such tests), and have medium stakes for educators (i.e., consequences such as loss of autonomy accrue indirectly and only over time). However, tests may have high stakes for some educators (e.g., North Carolina provides annual bonuses to teachers in schools meeting excellence criteria), and if the school fails to

make AYP for many consecutive years, the consequences for educators may be substantial (e.g., loss of pay or job). The phrase "high stakes test" is used in this article to refer to tests states use to implement AYP decisions under NCLB.

USING TEST DATA TO EVALUATE SCHOOLS

Although NCLB mandates the use of test data for determining whether schools and LEAs meet AYP targets, the data generated by annual testing may be used for other purposes. Different purposes often require different methods to analyze and present data. Currently, three broad models are recognized for evaluating schools or LEAs: (a) status models, (b) improvement models, and (c) growth models (Council of Chief State School Officers [CCSSO], 2005). Each of these models is described in the following sections.

Status Models

A "status model" of accountability sets a target for student performance at a given point in time, without regard to past or future performance. The primary definition of AYP in NCLB is a status model. That is, each state must measure students' achievement of standards, and each state must identify objective criteria to determine whether students are proficient for their grade level. "Proficiency" is a criterion-referenced judgment that varies by grade level (e.g., students in eighth grade must score higher on a mathematics test to be proficient than students in third grade). NCLB requires states to judge schools AYP on the basis of the proportion of students who score at or above this "proficient" level. Although start points and annual targets vary by state, and within states by subject matter and grade level, schools must ensure that a certain proportion of their students (within each of nine groups) meets or exceeds the target. All states must set the target for 2014 at 100%, meaning all states must require all students to attain 100% proficiency.

For example, in 2005-2006, schools in North Carolina must have at least 76.7% of students in grades 3-8 proficient in reading, and 81% proficient in mathematics. In 2007-2008, the targets increase to 84.4% and 87.3%; in 2010-2011, they jump to 92.2% and 93.7%. It does not matter whether there are changes in the student body, or how well the school has done in the past, nor whether students in one school start at a different place than students at another school. The status model simply

sets the target for a given year, then determines whether the target was met for the groups for which the school is responsible (which is determined by grade, demographic status, and group size).

The status model has the virtues of simplicity of calculation, ease of understanding, and defining and enforcing similar outcomes for all groups. However, the status model is insensitive to improvement in students and schools and is highly influenced by non-school factors (e.g., students' SES, parental education, ethnicity), which renders it a poor indicator of school quality (McCall, Kingsbury, & Olson, 2004; Raudenbush, 2004). Indeed, there is little evidence to support the argument that status models are even loosely associated with school quality (Haertel, 1999), and so the model is not widely supported by scholars, researchers, and organizations associated with assessing school quality (Linn, 2005).

Improvement Models

In contrast to status models, improvement models use the performance of students in a previous year or years, along with performance of students in the current year, to decide whether a school is making appropriate progress. This model is considered an improvement model because it may allow a school not meeting a status target to nonetheless meet AYP goals because a larger proportion of students in the current year are proficient compared to cohorts in previous years. The current "safe harbor" provision of NCLB outlines an improvement model alternative for schools to meet AYP. That is, if a school falls short of the status model target for a given group, if the proportion of students in that group scoring below proficient is reduced by 10% from the previous school year, and the group made progress on other academic indicators (e.g., attendance or graduation rate), the school may be considered to be making AYP.

The advantages of the improvement model include acknowledging that some schools have more challenging students than others, and that annual improvements in school performance should be recognized in an accountability system. The problems with the improvement model include setting different standards for different groups (e.g., why should a lower proportion of proficiency be accepted for a group even if it is better than the previous year's performance?) and comparing different cohorts across different years. The latter problem is particularly vexing for schools that experience changes in student demographics. For example, newly available subsidized housing, changes in the numbers of

students attending a school, or district reassignment of pupils to schools, could substantially change the proportion of students who are proficient across different years, yet improvement models do not consider such changes in deciding whether improvement occurred (CCSSO, 2005; McCall et al., 2004).

Growth Models

Growth models attempt to measure change (i.e., "growth") over one or more years within the same students attending the same school. Therefore, a student's change from third to fourth grade, or from sixth to eighth grade, might be used to evaluate school effectiveness. The argument posits that schools that produce greater rates of growth among their students are better than schools producing lower rates of growth.

Growth models may also consider nonschool factors in evaluating school performance. This is important, because students do not have the same rates of growth when compared across different demographic and ability groups (Meyer, 1996). For example, students from poor families have lower rates of growth than students from wealthy families; likewise, students who have more knowledge at the beginning of a school year tend to improve more than students with less knowledge. Growth models that statistically control for nonschool factors (e.g., student demographics, ability) in evaluating growth are often called "value added" models, because they attempt to remove such variables from judgments about the degree to which schools "add value" to students in successive years (see CCSSO, 2005, for a discussion).

Although intuitively appealing, growth models have a number of drawbacks. The most critical of these drawbacks include the statistical and logistical capacity to measure and model growth. Not only are growth calculations complex, but the ability to produce and use a continuous scale to measure student achievement across two or more years, set growth targets, and otherwise implement growth models is a significant challenge (CCSSO, 2005). Some tests may have scales that lend themselves to reflecting more rapid growth in moving from low to average performance, whereas other tests may be more sensitive to growth from average to high performance (see Ferrara, Johnson, & Chen, 2004; Lissitz & Huynh, 2003; and Reckase & Martineau, 2004, for measurement challenges). Additionally, growth models may be criticized for failing to define and enforce equal expectations across groups, particularly if value-added models are used that adjust for no-school effects (i.e., by adjusting for ethnicity and SES, one is essentially setting a dif-

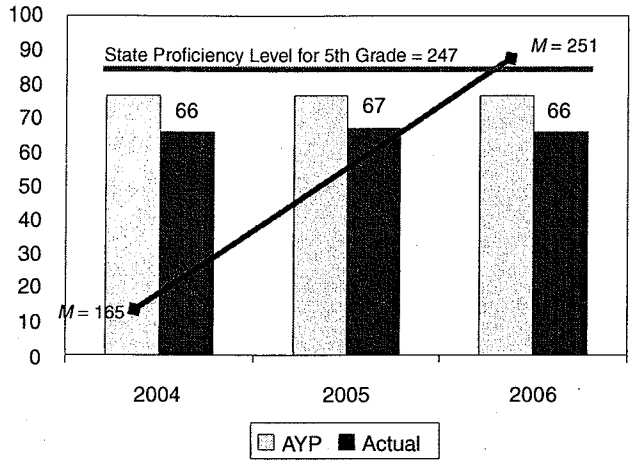
ferent standard for schools with high vs. low concentrations of ethnic/SES groups). A final limitation of growth models is that one must have multiple data points (ideally, three or more) to reliably predict and then measure growth. Given the costs of annual assessments, most states do not test in grades other than those required under NCLB (i.e., grades 3-8 and once in high school). To use NCLB data for growth modeling, a school should have three of these grades, meaning only elementary (grades 3-5) and middle schools (grades 6-8) are likely to be poised to use growth models for accountability and evaluation. Because most secondary schools only test students in one grade (usually tenth), or only in certain courses, multiple data points are not available for growth model evaluations of secondary schools.

Despite these problems, most scholars and educators agree that growth models are more fair than status or improvement models for measuring school quality, because growth models more directly determine which schools are doing better (or worse) than expected given the student bodies that they serve. That is, some schools that fail AYP under a status model actually improve student learning more than average, and some schools with excellent rates of proficiency do little to enhance student growth (McCall et al., 2004). Although NCLB does not currently allow for the use of growth models in determining AYP, the US Department of Education (2006) recently announced that two states, Tennessee and North Carolina, could begin using growth models as an alternative to status models for deciding whether a school meets AYP. Preliminary results in North Carolina using the growth model found that slightly less than 60 schools that failed to make AYP using the current status and improvement (i.e., "safe harbor") models would make AYP under the new growth model system. Although this is heartening news, it is a relatively small portion (about 6%) of the 968 schools that failed to make AYP in that year. Therefore, although growth models may produce different judgments of school quality than status models in some cases, it does not appear that a large proportion of schools failing under a status model are producing high or unexpected rates of growth within their students.

Status vs. Improvement vs. Growth Models: An Example

Figure 1 illustrates school data to illustrate the different kinds of conclusions one might draw from high stakes test data using status, improvement, and growth models. The figure shows reading test data for students with low income enrolled in a North Carolina elementary

FIGURE 1. Data Contrasting Status, Improvement, and Growth Models for a North Carolina School.



school in 2004, 2005, and 2006. The AYP annual targets for the school are a constant 76.7% each year (represented by the horizontal line). The school's percent proficient in grades 3-5 is represented in the solid dark vertical bar, and varies from 66-67%. Therefore, a status model would hold that the school did not meet AYP goals in any of the three years. Likewise, because the percent proficient across the three years does not show reliable or substantial increases, the school will not meet AYP targets using a "safe harbor" or improvement model.

The solid black line illustrates the mean scale score on the state test achieved by a cohort of students that were in the third grade (in 2004), and stayed with the school through fifth grade (in 2006). The state mandated level for proficiency in fifth grade is represented by the horizontal line (a scale score value of 247). It is clear that, in third grade, this cohort is well below fifth grade proficiency. However, by fifth grade, the cohort has progressed to the point where their average is above the state's proficiency cutoff, so most members score in the proficient range. The growth in this cohort implies that the school is highly effective at moving low scoring students to a level of proficiency by the time they complete fifth grade. A number of reasons might explain why the percent proficient in the years sampled does not change despite improvement in this cohort, such as student mobility or changes in student demograph-

ics (e.g., increasing numbers of low scoring students in third grade might offset gains in later grades).

In this example, it is clear that the model used to evaluate school performance leads to substantially different conclusions about the quality of the school. Status and improvement models identify the school as ineffective, whereas the growth model suggests the school is remarkably effective. Note that this example, although realistic, is not representative. Most of the schools with high and low status also exhibit high and low growth (McCall et al., 2004). However, some schools are characterized differently by the three models, with low status/high growth schools (such as the one in this example) being misrepresented as failing, and high status/low growth schools being misrepresented as successful. For a detailed comparison of status and growth models that uses high stakes test data to evaluate special education programs, see Schulte and Villock (2004).

Growth models also make it possible for schools to use data to evaluate the quality of the experiences they provide. In contrast to status and improvement models, which do not allow schools to determine what they add to student achievement, growth models could help schools identify which programs and practices within the school are most (and least) effective for promoting student growth. Issues of appropriateness, capacity, and consequences must be considered carefully in any use of test data. Therefore, the remainder of this article considers characteristics of test data that influence their value, and how to use such data to improve results.

CHARACTERISTICS THAT INFLUENCE THE VALUE OF HIGH STAKES TEST DATA

High stakes test data are similar to lights on an automobile dashboard: They alert the driver to a problem, but do not provide diagnostic information regarding what is causing the problem or what the driver should do about it. Drivers ignore such warnings at their peril, and good drivers will quickly seek additional information to diagnose and remedy the problem. Likewise, good educators will use annual test data to help identify problems, and will seek additional information to diagnose and remedy the problem.

There are four characteristics of test data that influence their value for program planning and evaluation: (1) breadth vs. narrowness of sampling, (2) timing of data availability, (3) the unit of analysis for aggrega-

tion of data, and (4) the metric in which data are reported. Each of these characteristics influences the value of test data for program planning and evaluation.

Breadth of sampling Most high stakes tests are designed to sample broad academic domains. Although broad representation of many skills increases the validity of tests as indicators of students' achievement of state standards, this same feature may limit value for program planning and evaluation. Knowing that a relatively high proportion of students are not proficient in "reading" is useful for identifying a problem, but not useful for understanding what to do about it. Deconstruction of broad domains into meaningfully inter-related units, such as the five components of reading (Armbruster, Lehr, & Osborn, 2001; National Reading Panel, 2000) or the seven domains of mathematics (National Council of Teachers of Mathematics, n.d.), is helpful in guiding instructional responses. However, high stakes tests may not have a sufficient number of items to adequately measure specific skills. For example, the proportion of students in Wisconsin considered proficient in Reading: Evaluation and Extend Meaning subskill went from 54% to 82% between the years of 2000 and 2001 (Department of Public Instruction, 2002) simply because one of the items used to estimate performance on that objective went from being a relatively easy to a relatively difficult item. Therefore, school psychologists should consider carefully issues of breadth, narrowness, and the influence of item changes in representing performance in academic objectives and skills.

Timing of test data. Two features of timing influence the uses and value of test data for program planning and evaluation. The first feature is the frequency with which data can be made available to those who need data to make programmatic and instructional decisions, and the second feature is the point at which data become available to those who use the data. Generally, higher frequency is better than lower frequency (e.g., Shinn, 2002), and formative data (i.e., data that can be used for planning) are better than summative data (i.e., those that only determine the degree to which an outcome has been achieved). Because high stakes test data are sampled annually, and usually are not available until the end of the academic year, they are of little or no value for program planning for test takers. However, their use for strategic planning (i.e., planning for the following academic year) will be discussed in the last section of this article.

Unit of analysis. High stakes test data are, by law, reported at the individual, group, school, and district level. However, these units do not necessarily correspond to the units that matter for program planning and

evaluation. Rather, instructional groupings are more relevant. There are two ways to align data to instructional groups. The first is to aggregate test data to match existing structures (e.g., classrooms, individuals enrolled in a program). The second is to group students by test score data (i.e., assign students to groups based on test data). There are risks associated with both approaches; in the former, one risks unintended consequences (e.g., divisive comparisons between classrooms, teachers, and programs), whereas in the latter, one risks self-fulfilling prophecies with respect to low, medium, and high performing students.

Metric for reporting data. One of the more interesting features of NCLB is the requirement to report frequencies and proportions of students in a given group who score Proficient or Advanced on the state test. This requirement has two implications for metrics in which results are reported. The first implication is that tests must measure student performance on a criterion-referenced scale, rather than a norm-referenced scale (e.g., it is possible for the majority of students to be above proficient, but not above average). The second implication is that measures of central tendency, such as a mean or median, are irrelevant. Therefore, exceptionally high or low test scores mean the same as those just over or under the cutoff, which makes exceptionally high (or low) achievement irrelevant to AYP decisions. This may cause cynical educators to shift resources away from exceptional (i.e., high- or low-scoring) students in favor of students who score slightly below proficient to improve AYP performance.

The four features of high stakes test data (i.e., breadth, timing, unit of analysis, and reporting metric) influence the ways in which such data should–and should not–be used for program planning and evaluation. The following section attempts to outline practices that are intended to enhance the ability of school psychologists to plan and evaluate programs with the goal of enhancing students' academic performance.

USING RESULTS TO IMPROVE RESULTS

Given the limitations of high stakes test data, how can school psychologists use these data to enhance student success? The answer to this question is not well supported by research. However, research on effective schools (e.g., Newmann & Wehlage, 1995) and field studies of improving schools (e.g., Porter, 2002) suggests alignment of instruction and assessment is likely to improve outcomes. That is, schools that clearly identify what they want students to know and do, and then align

instruction to ensure adequate opportunity for students to acquire such knowledge and processes, are more effective than schools that do not align their effort toward clearly identified goals. The standards-based reform movement was largely intended to help schools better align instruction to state standards (Swanson, 2006). Table 1 provides resources that may help psychologists use high stakes test data to improve educational outcomes.

Figure 2 provides a decision process for using high stakes test data to enhance instructional opportunities for students. The decision tree presumes that data are typically available for groups of students that make instructional sense (e.g., grades, classrooms, or groups targeted by a particular instructional program). Furthermore, decisions presume that data are available at or near the end of the school year, meaning that the purpose of the decisions is to inform instructional planning for the following year's cohort. Finally, decisions presume that those who will be executing instructional or programmatic changes are present and engaged in the process outlined in the tree. The school psychologist poses the questions and facilitates decisions, but the decisions should be made by those who are responsible for implementation.

Are Students Meeting AYP in the Subject Matter Area?

The answer to this question is fairly straightforward, and assumes failure to meet AYP is due to performance (i.e., score-related) criteria, rather than participation criteria. Should a school fail to make AYP because of missed participation goals, the focus of the response should be towards more effective inclusion of students and subgroups in the general curriculum (McDonnell, McLaughlin, & Morrison, 1997) and state tests (Elliott, Braden, & White, 2001). If the school fails to make AYP performance goals, one must ask if this is a general problem (i.e., Do many students fail to make AYP?) or a specific (i.e., only one or more subgroup) problem?

If most groups are making AYP, and some are not, a school team may elect to target "at risk" students. Although targeting allows schools to focus their limited resources on those most in need, targeting may also restrict opportunities to learn (e.g., targeted students may have less opportunity to learn social studies because their reading time is increased). More widespread failure argues in favor of more general, rather than targeted, school changes; conversely, meeting AYP targets tends to validate current practices and programs.

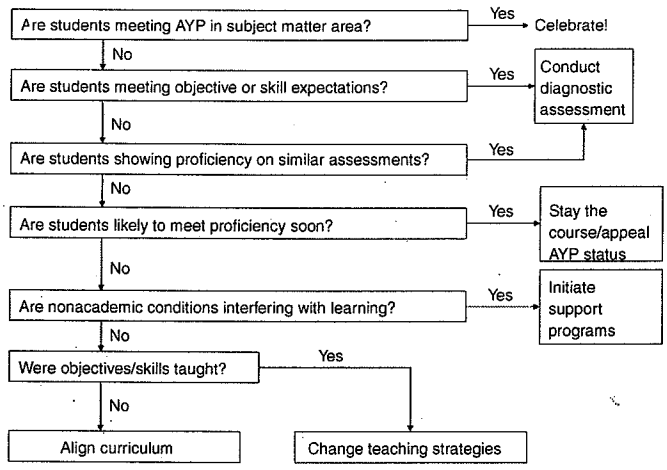
TABLE 1. Resources to Guide Selection and Implementation of Evidence-Based Practices to Improve Student Outcomes

Resource	URL
Assessing Opportunity to Learn in Schools (Stevens, 1999)	*http://www.temple.edu/LSS/htmlpublications/spotlights/300/spot307.htm*
Assessing Opportunity to Learn Survey (Stevens, 1999)	*http://www.temple.edu/lss/pdf/publications/pubs1999-8appendix.pdf*
Considerations when Selecting a Reading Program	*http://www.k8accesscenter.org/training_resources/readprograms.asp*
CRESST (Center for Research on Evaluation, Standards, & Student Testing)	*www.cresst.org*
Defining, Developing, and Using Curriculum Indicators (Porter & Smithson, 2001)	*http://www.cpre.org/Publications/rr48.pdf*
Educators' Guide to Schoolwide Reform (Am. Institutes for Research)	*http://www.aasa.org/issues_and_insights/district_organization/Reform/*
English Learners: Assessing Opportunity to Learn (OTL) in Grade 6 Language Arts(Boscardin, et al., 2004)	*http://www.cse.ucla.edu/r/l.asp?r=728*
ERIC Digest on Opportunity to Learn (Schwartz, 1995)	*http://www.ericdigests.org/1996-3/urban.htm*
Evidence-based interventions searchable data base (Campbell Collaborative)	*http://www.campbellcollaboration.org/*
Improving student performance in math (US Dept. of Ed.)	*http://www.ed.gov/teachers/how/math/edpicks.jhtml?src=ln*
Improving student performance in reading (US Dept. of Ed.)	*http://www.ed.gov/teachers/how/read/edpicks.jhtml?src=ln*
Issues in Assessing English Language Learners' Opportunity to Learn Mathematics (Herman & Abedi, 2004)	*http://www.cse.ucla.edu/r/l.asp?r=726*
Mathematics Instruction Resources (National Academy Press)	*http://www.nap.edu/catalog/10434.html http://www.nap.edu/catalog/9822.html*
National Institute of Child Health and Human Development (NICHD) Reading research & practice resources	*http://www.nichd.nih.gov/crmc/cdb/reading.htm*
National Institute for Literacy (research, interventions, publications preK-adult)	*http://www.nifl.gov/*
National Reading Panel	*www.nationalreadingpanel.org*
NCREL Opportunity to Learn site	*http://www.ncrel.org/sdrs/areas/issues/methods/assment/as8lk18.htm*
Oregon's Reading First Review of Curricula/Programs (Science/materials alignment)	*http://reading.uoregon.edu/curricula/or_rfc_review_2.php*

TABLE 1 (continued)

Resource	URL
Reading intervention/prevention resources (National Academy Press)	*http://books.nap.edu/catalog/10130.html http://books.nap.edu/catalog/6023.html http://books.nap.edu/catalog/6014.html*
Rigorous research policy statement by Robert Slavin	*http://www.americanprogress.org/site/pp.asp?c=biJRJ8OVF&b=492641*
Standards for interventions based on rigorous evidence (US Dept. of Ed.)	*http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html*
Survey of Enacted Curriculum	*www.SECsurvey.org*
Task Force for Evidence-Based Interventions (APA Div. 16, SSSP)	*http://www.indiana.edu/~futures/kratochwill.pdf*
Ten Myths of Reading Instruction, Southwest Educational Development Laboratory	*http://www.sedl.org/pubs/sedl-letter/v14n03/2.html*
What Works Clearinghouse	*www.w-w-c.org*
Writing Difficulties Prevention and Intervention for Students w/ LD (Graham, Harris, & Larsen, 2001)	*http://www.ldonline.org/ld_indepth/writing/prevention_intervention.html*
Writing interventions (meta analysis by Gersten, Baker, & Edwards)	*http://www.ld.org/research/ncld_writing.cfm*

FIGURE 2. A Decision Tree for Using High Stakes Test Data for Program Planning and Evaluation



Are Students Meeting Objective or Skill Expectations?

AYP judgments are based on composite (i.e., broad) scores in reading, mathematics, and science. Many states also provide information on specific objectives or skills within an academic domain. Analysis of students' performance on objectives may help pinpoint causes of general academic deficits. Ideally, objective scores should be reported as a proportion of the student group meeting or exceeding proficiency, but some states report objective performance using means. If a school is working with means, it is useful to consider if the distribution of scores are approximately normal (indicating a uniform group of students). Non-normal distributions imply different responses. For example, positively skewed distributions (where most students do poorly, but the mean is inflated by some exceptionally high scorers) underestimates the magnitude of the problem, whereas negatively skewed distributions (i.e., most students do well, but the mean is deflated by some exceptionally low scorers) over-estimate the problem. Multimodal distributions imply two or more distinctly different groups. This pattern suggests targeting some students but not others.

The goal of this step is to develop hypotheses about instruction and programming that might account for failure to meet AYP. Generally, poor performance on one or more objective suggests deficits in how those objectives are taught. Additional diagnostic assessment of a small (ideally, randomly selected) group of students may help identify specific skill deficits within an academic domain. However, keep in mind that objective skill performances are estimated with a small number of items, and are therefore unreliable, which leads to the next step in the process.

Are Students Showing Proficiency on Similar Assessments?

The preceding step should generate hypotheses about problems with current instruction and curricula; this stage should test those hypotheses by obtaining additional data. There are two reasons why school psychologists must collect additional data to evaluate hypotheses generated from high stakes test data. The first is scientific integrity; the same data that lead to a hypothesis cannot be used to test the hypothesis. The second reason is to enhance stakeholder understanding of and commitment to the hypothesis. Many educators question the value of high stakes test data (Johnson, Arumi, & Ott, 2006), but if they reach the same conclu-

sion about causes of poor student performance by data that they value, they will be more likely to accept the hypothesis as valid and useful for considering instructional, curricular, and programmatic changes.

Independent data may come from other tests (if available); analysis of student work samples, in-class quizzes, and exams; and teachers' observations of student performances. Teachers are generally good at predicting how students will do on specific tasks or test items (Demaray & Elliott, 1998), so it is important to provide teachers with clear examples of the kinds of tasks and the degree of proficiency demanded by high stakes tests. When provided with such structure, teachers are often quite accurate in estimating student skills, although they tend to underestimate skills for students with disabilities (Hurwitz, Elliott, & Braden, in press). If additional data are required, sampling methodology (e.g., groups of 20 students) and targeted assessments (e.g., criterion-referenced tasks that may be administered in a few minutes) can produce them quickly and relatively cheaply. Ideally, stakeholders should collect data, and may use nonrandom sampling (e.g., selecting 10 students who passed and 10 who failed a particular objective) to evaluate the veracity of high stakes test results.

Are Students Likely to Meet Proficiency Soon?

Schools and LEAs that have strong progress monitoring systems as part of the general education program will not only have extant data to help test hypotheses in the previous stage, but may also be able to predict rates of growth towards AYP proficiency goals. Progress demonstrated on curriculum-based measures and other progress monitoring tools such as DIBELS (Good, Gruba, & Kaminski, 2002) are good to excellent predictors of who will and will not meet proficiency on high stakes tests (Ax, 2004; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006). If progress monitoring data, or longitudinal data on high stakes tests (i.e., growth models) indicate that low-scoring students are likely to acquire proficiency, it could lead the team to decide to "stay the course" with respect to current instruction, curricula, or programs. Note, however, that such judgments require strong progress monitoring systems (see the National Center on Student Progress Monitoring, n.d.) and strong measurement and technical capacities to reliably identify trends towards proficiency. Such predictions should not be based on teacher judgments.

Are Nonacademic Conditions Interfering with Learning?

Although AYP and high stakes tests focus on students' academic proficiency, nonacademic indicators such as student attendance, disciplinary events, psychosocial screenings, and teacher observations may indicate nonacademic causes for low academic proficiency. School psychologists should be skeptical about accepting nonacademic causes, because many off-task behaviors are motivated by students' seeking to escape inappropriate instructional demands (Hanley, Iwata, & McCord, 2003). Coordinating academic and nonacademic responses is usually better than exclusively focusing on nonacademic features to improve student learning. Resources to evaluate and intervene with nonacademic features of programming include the Research and Training Center for Children's Mental Health (see Kutash, Duchnowski, & Lynn, 2006), the UCLA Mental Health Project Center for Mental Health in Schools (e.g., Center for Mental Health in Schools, 2003), and the National Technical Assistance Center on Positive Behavioral Interventions and Supports (2004).

Were Objectives/Skills Taught?

The presumptive answer to this question is "Yes"; however, careful inspection of teaching behaviors and conditions may suggest students lacked sufficient opportunities to learn. There are many approaches to identifying opportunities to learn (Schwartz, 1995), including structured self-report surveys, direct observations, teacher journals, student surveys, and teacher interviews (see Porter, 2002). Surveys of the "enacted curriculum" (Council of Chief State School Officers, n. d.; Porter & Smithson, 2001) list state academic objectives and invite teachers to report the degree to which they offer frequency and depth on instruction for each objective. Teachers' instructional allocations strongly predict students' performance on high stakes tests (Herman & Abedi, 2004). Less formal approaches to evaluate students' opportunities to learn include analysis of lesson plans, examination of work assigned to students, and teacher absenteeism.

Deciding What to Do: Alignment vs. Change

NCLB presumes that schools afford students opportunities to learn the standards adopted by each state, and that those opportunities should be guided by scientific, research-based methods of instruction. The

practice of ensuring that schools afford students opportunities to learn state standards is termed alignment, meaning the school aligns its instructional activities and materials to state standards. Determination that the curricula provided and the standards assessed in high stakes tests are not aligned suggests increasing alignment as a logical focus of program planning, whereas determining that curricula are aligned–but students have not acquired proficiency–suggests changes in the nature of learning opportunities, such as methods and materials used in instruction, should be the focus of program planning.

Alignment. When the decision process indicates that students have not had sufficient opportunities to learn standards, program stakeholders should align program activities to state standards. Although there are many approaches to increasing alignment, some of the most common are:

Vertical integration of curricula across grade levels. Often, what is taught at one grade is not well coordinated with what is taught at other grades. When curricula are not thoughtfully integrated between grades, some aspects of the curriculum may receive more attention than they need, whereas others get less.

Annual lesson planning within a grade. Although flexibility and adaptability are essential characteristics of any teacher, some standards may be neglected because teachers do not manage learning opportunities across the academic year. Careful consideration of what needs to be covered when (i.e., scope and sequence) must be balanced with non-academic aspects of the school year (e.g., the time between Thanksgiving and winter holiday breaks is better for review than introduction of new information).

Selection of teaching materials. The alignment between instructional materials (e.g., textbooks) and standards influences learning opportunities. Therefore, educators must consider using supplemental materials (or omitting portions of a text) when developing annual lesson plans.

In guiding this decision-making process, school psychologists should recognize that educators tend to assign meanings to state standards that match their current expectations and contexts, rather than appreciate the degree to which standards might diverge from their current expectations and contexts (Hill, 2001; Ogawa, Sandholtz, Martinez-Flores & Scribner, 2003).

Change programming. If the program evaluation process indicates that students have been afforded appropriate opportunities to learn, but still do not exhibit proficiency on high stakes tests, the school or program should consider whether current practices could be replaced by

more effective and efficient (i.e., evidence-based) methods to support student learning. Adoption of evidence-based practices is particularly important for schools serving students who are at risk of or currently experiencing academic failure (Slavin, 2005). However, the ability of practitioners to reliably identify and implement evidence-based practices is constrained, in part because there is only limited availability of vetted sources, and in part because collecting, reading, rating, and categorizing the literature on a given practice is an extremely time-consuming process that invokes disagreements even among highly trained professionals. Until reputable sources (e.g., the What Works Clearinghouse) provide a list of vetted practices, practitioners may find it quite difficult to identify, much less implement, evidence-based practices at their schools.

The resources needed to implement and sustain changes in programs or instruction must be considered when replacing current practices with evidence-based practices. Underestimation of the resources needed to implement substantial change may undermine the sustainability of long-term improvements. School psychologists should also consider carefully the unintended consequences of test use, and take steps to minimize those consequences in planning their activities (see Jones, 2007; Kruger, Wandle, & Struzziero, 2007).

CONCLUDING COMMENTS

This article is intended to help school psychologists understand and use high stakes test data to evaluate and improve educational programs. The proposed decision-making process attempts to capitalize on assets unique to school psychologists, including knowledge of assessment, ability to understand test results, quantitative skills for analyzing and presenting test data, and strong grounding in evidence-based practices. However, school psychologists are unlikely to be in a position of administrative authority over educational stakeholders, and so the role herein described is one of data-based consultation. It must be acknowledged that, although such a role is consistent with professional practice guidelines (e.g., Ysseldyke, Burns, Dawson, Kelly, Morrison, Ortiz, Rosenfield, & Telzrow, 2006), it is not one that would meet the standards identified for evidence-based practice–that is, there is no rigorous experimental evidence to show that such a role reliably enhances student outcomes. Therefore, school psychologists are cautioned to be critical consumers of this (and other) recommendations for practice; and to con-

sider carefully the consequences of inaction in response to the increasing pressures created by high stakes tests, and the consequences of choosing not to participate in the presentation, understanding, and use of high stakes test data for program planning and improvement.

REFERENCES

Armbruster, B. B., Lehr, F., & Osborn, J. (2001). *Put reading first: The research building blocks for teaching children to read.* Washington, DC: US Department of Education.

Ax, E. E. (2004). *Relationship between curriculum-based measurement reading and statewide achievement test mastery for third grade students.* Unpublished Masters thesis. University of South Florida, Tampa, FL. Retrieved July 6, 2006, from: http://purl.fcla.edu/fcla/etd/SFE0000568.

Braden, J. P. (2002). Educational accountability: High stakes testing and educational reform. In A. Thomas & J. Grimes. (Eds.): *Best practices in school psychology (4th ed.)*, pp. 301-319. Silver Spring, MD: National Association of School Psychologists.

Braden, J. P., & Tayrose, M. P. (in press). Best practices in educational accountability: High stakes testing and educational reform. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology (5th ed.)*. Silver Spring, MD: National Association of School Psychologists.

Center for Mental Health in Schools (2003). *An introductory packet on assessing to address barriers to learning.* Los Angeles, CA: Author. Retrieved 3 July, 2006, from http://smhp.psych.ucla.edu/pdfdocs/barriers/barriers.pdf.

Center on Positive Behavioral Interventions and Supports (2004). *School-wide positive behavior support: Implementers' blueprint and self-assessment.* Eugene, OR: Author. Retrieved June 13, 2006, from http://www.pbis.org/files/Blueprint%20draft%20v3%209-13-04.doc.

Council of Chief State School Officers (no date). *Surveys of enacted curriculum.* Washington, DC: Author. Retrieved 18 June, 2006, from www.SECsurvey.org.

Council of Chief State School Officers (2005, October). *Policymakers' guide to growth models for school accountability: How do accountability models differ?* Washington, DC: Author. Retrieved 18 June, 2006, from http://www.ccsso.org/content/pdfs/Growth%20Models%20Policymaker%20 Guide%202005.pdf.

Demaray, M., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*(1), 8-24.

Department of Public Instruction (2002, August). *Objectives performance summary* (data file). Available from Wisconsin Department of Instruction web site, http://dpi.wi.gov/oea/xls/opi99-02s.xls.

Education Week. (2006). Quality Counts at 10: A decade of standards-based education. *Education Week, 25*(17). (Retrieved June 18, 2006, from http://www.edweek.org/ew/toc/2006/01/05/).

Elliott. S. N., Braden, J. P., & White, J. L. (2001). *Assessing one and all: Educational accountability for students with disabilities*. Reston, VA: Council for Exceptional Children.

Ferrara, S., Johnson, E., & Chen, W. H. L. (2004, April). *Vertically moderated standards: Logic, procedures, and likely classification accuracy of judgmentally articulated performance standards*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego.

Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes. (Eds.): *Best practices in school psychology (4th ed.)*, pp. 699-720. Silver Spring, MD: National Association of School Psychologists.

Haertel, E. H. (1999). Validity arguments for high stakes testing: In search of the evidence. *Educational Measurement Issues and Practice, 18*(4), 5-9.

Hanley, G., Iwata, B., & McCord, B. (2003). Functional analysis of problem behavior: A review. *Journal of Applied Behavior Analysis, 36*(2), 147-185.

Herman, J. L., & Abedi, J. (2004). *Issues in assessing English Language Learners' opportunity to learn mathematics* (Center for the Study of Evaluation Report No. 633). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Available: www.cresst.org

Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Retrieved, January 4, 2006, from http://books.nap.edu/execsumm_pdf/6336.pdf.

Hill, H.C. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal, 38*(2), 289-318.

Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (in press). The influence of test familiarity and student disability status upon teachers' judgments of students' test performance. *School Psychology Quarterly*.

Johnson, J., Arumi, A. M., & Ott, A. (2006). *Is support for standards and tests fading?* (Reality Check Educational Insights Issue No. 3). Washington, DC: Public Agenda. Retrieved June 18, 2006, from (http://www.publicagenda.org/research/research_reports_details.cfm?list = 100).

Jones, B. (2007). The unintended outcomes of high stakes testing. *Journal of Applied School Psychology, 23* (2), 67-88.

Kruger, L. J., Wandle, C., & Struzziero, J. (2007). Coping with the stress of high stakes testing. *Journal of Applied School Psychology, 23* (2), 109-128.

Kutash, K., Duchnowski, A. J., & Lynn, N. (2006). *School-based mental health: An empirical guide for decision-makers*. Tampa, FL: The Research and Training Center for Children's Mental Health, Louis de la Parte, Florida Mental Health Institute, University of South Florida. Retrieved July 3, 2006, from: http://rtckids.fmhi.usf.edu/rtcpubs/study04/.

Linn, R. L. (2005). *Fixing the NCLB accountability system* (CRESST Policy Brief No. 8). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Lissitz, R. W. & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability.

Practical Assessment, Research, & Evaluation, 8(10). Retrieved March 30, 2006 from http://PAREonline.net/getvn.asap?v = 88&n = 10.

McCall, M. S., Kingsbury, G. C., & Olson, A. (2004). *Individual growth and school success.* Lake Oswego, OR: Northwest Evaluation Association. Retrieved 24 June, 2006, from http://www.coloradoleague.org/New%2520Leadership%2520Wkshp%252004/NWEA%2520Growth.pdf.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (Eds) (1997). *Educating one & all: Students with disabilities and standards-based reform.* Washington, DC: National Academy Press.

Meyer, R. L. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 171-196). Washington, DC: National Academy Press.

National Center on Student Progress Monitoring (no date). *What is progress monitoring?* Washington, DC: Author. Retrieved 10 July, 2006, from http://www.studentprogress.org/.

National Council of Teachers of Mathematics (no date). *Standards for school mathematics.* Reston, VA: Author. Retrieved July 3, 2006, from http://www.nctm.org/standards/standards.htm.

National Reading Panel (2000). Report of the National Reading Panel: Teaching children to read–An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (Summary). Washington, DC: US Department of Health and Human Services.

Newmann, F. M. & Wehlage, G. G. (1995). *Successful school restructuring: A report to the public and educators.* Madison, WI: Center on Organization and Restructuring of Schools, Wisconsin Center for Education Research, University of Wisconsin-Madison.

Ogawa, R.T., Sandholtz, J. H., Martinez-Flores, M., & Scribner, S. P. (2003). The substantive and symbolic consequences of a district's standards-based curriculum. *American Educational Research Journal, 40*(1), 147-176.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*, 3-14.

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators* (Research Report Series RR-048). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education. Retrieved June 18, 2006, from http://www.cpre.org/Publications/rr48.pdf.

Raudenbush, S. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* (Report PIC-ANG9). Princeton, NJ: Educational Testing Service. Retrieved June 18, 2006, from http://www.ets.org/portal/site/ets/menuitem.c988ba0e5dd572bada20bc47c3921509/?vgnextoid=4fffaf5e44df4010VgnVCM10000022f95190RCRD&vgnextchannel=e2a5be3a864f4010VgnVCM10000022f95190 RCRD.

Reckase, M. D. & Marineau, J. (2004, Oct.). *The vertical scaling of science achievement tests.* National Academy of Sciences. Retrieved 12 June, 2006, from http://www7. nationalacademies.org/bota/Vertical%2520Scaling.pdf.

Schulte, A. C., & Villock, D. N. (2004). Using high-stakes tests to derive school-level measures of special education efficacy. *Exceptionality, 12*(2), 107-126.

Schwartz, W. (1995). *Opportunity to learn standards: Their impact on urban students.* (ERIC/CUE Digest No. 110; ERIC Document No. 389816).

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24*(1), 19-35.

Shinn, M. R. (2002). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas & J. Grimes. (Eds.). *Best practices in school psychology (4th ed.)*, pp. 671-698. Silver Spring, MD: National Association of School Psychologists.

Slavin, R. E. (2005). *Evidence-based reform: Advancing the education of students at risk.* Washington, DC: Center for American Progress. Retrieved 3 July, 2006, from http://www.americanprogress.org/atf/cf/{E9245FE4-9A2B-43C7-A521-5D6FF2E06E03}/Slavin%203%2017%20FINAL.pdf.

Swanson, C. (2006, January). *Making the connection: A decade of standards-based education reform and achievement* (Editorial Projects in Education Research Center). Washington, DC: Education Week. Retrieved June 18, 2006, from http://www.edweek.org/media/ew/qc/2006/MakingtheConnection.pdf.

U.S. Department of Education (2006, May). *Growth models: Ensuring grade-level proficiency for all students by 2014.* Retrieved June 18, 2006, from http://www.ed.gov/admins/lead/account/growthmodel/proficiency.html.

Ysseldyke, J., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., Rosenfield, S., & Telzrow, C. (2006). *School psychology: A blueprint for training and practice III.* Bethesda, MD: National Association of School Psychologists.

doi:10.1300/J370v23n02_08