Comprehensive Accountability Systems
A Framework for Evaluation

Mid-continent Research for Education and Learning (McREL) is a private nonprofit corporation located in Denver, Colorado. We provide field tested, research-based products and services in the following areas:

- Assessment, Accountability, and Data Use
- Curriculum
- Diversity
- Early Childhood Education
- Education Technology
- Instruction
- Leadership and Organization Development
- Literacy
- Mathematics
- Professional Development
- Rural Education
- School Improvement and Reform
- Science
- Standards
- Teacher Preparation and Retention

Copyright Information

Comprehensive Accountability Systems
A Framework for Evaluation

Regional Educational Laboratory
Contract #ED-01-CO-0006
Deliverable #2002-07

Office of Educational Research and Improvement
U.S. Department of Education
Washington, D.C. 20208

prepared by

Bryan Goodwin, M.A., Senior Consultant
Kerry Englert, Ph.D, Senior Researcher
Louis F. Cicchinelli, Ph.D., Deputy Director

Mid-continent Research for Education and Learning
2550 S. Parker Road, Suite 500
Aurora, CO 80014
303-337-0990 (phone)
303-337-3005 (fax)
www.mcrel.org

Revised February 2003

This publication is based on work sponsored wholly, or in part, by the Institute of Education Sciences, U.S. Department of Education, under Contract No. ED-01-CO-0006. The content of this publication does not necessarily reflect the views of IES, the Department, or any other agency of the U.S. government.

TABLE OF CONTENTS

INTRODUCTION

BACKGROUND

During the 2000 presidential debates, one issue on which contenders George W. Bush and Al Gore found common ground was standards-based school accountability; during one televised debate Gore said, "Now, accountability? We basically agree on accountability."

Given the bipartisan support for school accountability, it's not surprising that it became one of the most prominent pieces of the No Child Left Behind Act of 2001. The Act requires all of the nation's publicly funded schools to make adequate progress toward demonstrating that 100 percent of their students are testing at proficient levels on statewide exams. Schools that fail to do so face severe consequences, including reconstitution, state takeover, or closure.

Proponents of this new form of tough, "no excuses" accountability say it's a concept that's long overdue. Setting clear expectations for all students and holding schools responsible for ensuring that all students meet those expectations will have several positive effects, including the following:

- Focusing schools and districts on learning outcomes, not process or input requirements
- Focusing teachers on helping *all* students achieve high levels of learning
- Providing schools with data they can use to make improvements in curricula or staffing
- Providing parents with more information about how their children and their children's schools are performing
- Providing the education community with needed data to determine what's working and what's not.

At the same time, however, many educators and researchers have expressed concerns about poorly constructed accountability systems creating a number of negative consequences, including encouraging schools and educators to

- narrow the curriculum;
- focus on bringing students to proficient levels, but not to advanced levels;
- increase retention rates, place more students in special needs, or provide inappropriate accommodations;
- focus more on responding to bureaucratic regulations rather than addressing other issues of greater concern to parents and the public (e.g., discipline, safety, values, workplace preparation); and
- adopt a compliance mentality, rather than a creative improvement mindset.

Both sets of arguments have merit. Standards-based accountability systems show too much promise to improve schools to be discarded altogether. As Scheurich, Skrla, and Johnson (2000) argue, "however flawed and imperfect these systems are, however much they need to be improved, three highly significant and even historic possibilities have emerged as a result of these systems" (p. 294). These possibilities include (1) the "high level of public attention paid to the lack of success of our public educational system with low-income children and children of color," (2) "the major public commitment by both political parties and their presidential candidates . . . to high academic performance for all races and socioeconomic classes of students," and (3) "the substantially improved academic success of children of color and low-income students and substantially improved equity in some schools and districts in some states" (pp. 294–295).

Although Scheurich et al. emphasize that these improvements have been made in "*some* schools and districts in *some* states" (p. 295), their point is that accountability systems have guided and encouraged real improvements in some locations. Therefore, rather than flatly rejecting or discarding such systems altogether, educators and policymakers should work to improve accountability systems so that they truly encourage positive outcomes. At the same time, the negative consequences created by some of these systems should be honestly examined and corrected.

As most policymakers acknowledge, outcomes-based accountability systems are complex — and relatively new — endeavors. As a result, most people responsible for designing existing accountability systems readily admit that these systems are still "works in progress." As such, it's reasonable to assume that they may need to be improved and modified to ensure that they are helping schools provide all students with quality opportunities for learning. Indeed, WestEd researchers have noted that a "critical element" of effective accountability systems is "a periodic, systematic checking on the impact of such a system," which asks the following kinds of questions:

> Is it [the accountability system] meeting its objective, namely, raising student and school performance? Have there been unintended consequences? What changes to the system need to happen to ensure its ongoing effectiveness? (Guth et al., 1999)

Sirotnik and Kimball (1999) offer the following rationale for proposing a set of "Standards for Standards-based Accountability Systems" in the pages of the *Phi Delta Kappan*:

> Accountability is the name of the game today and for the foreseeable future. So let's play by some rules. And let's construct those rules in good faith. After all, the public has a right to know how well the public schools are doing, just as the public schools have a right to be judged in reasonable and conscientious ways. If our students and our schools are to be held accountable for meeting "world-class" content and performance standards, then let's ensure that sensible standards exist against which the accountability systems themselves can be held accountable. (p. 211)

In short, as Sirotnik and Kimball argue, "holding accountability systems accountable is only reasonable" (p. 211). Only through continual evaluation of accountability systems can states, districts, and schools understand the negative or unintended consequences of such systems.

PURPOSE OF THIS REPORT

The purpose of this report is to provide a theoretical basis for a framework and process for evaluating accountability systems that McREL will develop in 2003. This process will be aimed at helping educators and policymakers evaluate their own accountability systems by

- re-examining and reflecting on the original purposes for these systems, and
- examining the extent to which their existing systems meet the characteristics of effective accountability systems based on current research findings.

After a brief discussion of the purposes of accountability systems, we offer a synthesis of existing recommendations and guidelines for accountability systems. We then provide three models for conceptualizing accountability systems that

- represent alignment of assessment and accountability system goals to allow for accurate and meaningful interpretations;
- provide fair, reliable, and valid data that support the use and interpretation of the results;
- provide the right mix of incentives and sanctions to stimulate changes that improve the learning outcomes of all students;
- provide information that educators and policymakers can use to guide school improvement efforts;
- restore the credibility of the system; and
- encourage parent and community engagement.

In the conclusion of this report, we provide some observations about what appears to be missing from the current dialogue surrounding accountability systems — that is, what questions are not being asked and what alternatives are not being considered. In 2003, we plan to integrate the discussions in this report about the purposes of accountability, existing recommendations and guidelines, conceptual models for accountability systems, and "big picture" issues to consider into a process that helps policymakers and educators continually review and improve their accountability systems.

ACCOUNTABILITY SYSTEM PURPOSES

Linn (2001) has noted that "the first question that should be asked of any assessment and accountability system is what are the purposes of the system?" (p. 2). Given that it has been several years since most accountability systems were first designed, it may be time to step back for a moment and review the intended purposes of accountability systems, and determine if those

systems are fulfilling their purposes. In doing so, states should look beyond the requirements presented in the recent federal legislation. That is, they should not see the federal requirements as the finish line for accountability, but rather the starting line.

Generally speaking, the most commonly cited purpose for accountability systems is to improve student learning. However, as Linn (2001) has pointed out, exactly how school accountability will lead to improvements in student learning is seldom specified. So it's worth examining the assumptions underlying this premise that accountability systems will lead to improvements in teaching and learning. The most common assumptions about outcome-oriented accountability systems appear to be that they will improve schools by

- informing students, parents, and teachers about student progress;
- monitoring the learning process and holding students, schools, educators, and states responsible for attaining learning outcomes;
- certifying teacher quality on the basis of student achievement;
- evaluating the overall effectiveness of schools or reforms and assisting education policymakers and administrators with programmatic decisions; and
- ensuring that equitable opportunities to learn are available for students.

Inform students, parents, and teachers about student progress. Publicly reporting assessment results can provide families with objective measures to help them understand how well their children are doing and what progress they are making over time. One assumption is that parents could use the information to demand educational improvement (Linn, 2001). In addition, the data can help educators determine the strengths and weaknesses of a student or group of students, and what skills and competencies individual students or entire groups of students need to work on.

Monitor the learning process and hold students, schools, educators, and states responsible for attaining learning outcomes. As Furhman (1999) and others have noted, states traditionally have monitored "compliance with input and process standards." (p. 1). The new focus is on outcomes, namely, student achievement. This focus is intended to "free schools from the old compliance mentality and to provide more flexibility so they can maximize student performance" (p. 1). Simply put, instead of asking how many books are in the library or how many children are coming to school, educators should be paying more attention to the following kinds of questions:

- Do all students have opportunities to learn?
- What should students know and be able to do?
- What is being taught?
- How are teachers teaching?
- Are our instructional programs effective?

Certify teacher quality on the basis of student achievement. Similarly, most would say that accountability systems are intended to create new ways to examine teacher quality. Rather

than simply defining teacher quality in terms of experience, education level, or seniority, the new systems provide a mechanism for defining teacher quality in terms of efficacy, or student outcomes. In so doing, they provide school leaders with information they can use to make better decisions about teacher professional development, placement, recruitment, and retention.

Evaluate the effectiveness of schools or reforms and assist policymakers and educators with programmatic decisions. Education historians, including Diane Ravitch (2000) and Larry Cuban (1990), have viewed the past decades of education reform with dismay, noting the tendency to implement new "fads," rather than "tried and tested" reforms. As a result, in recent years, there has been much discussion about the need to make education more scientifically based. By providing schools with better data and incentives to make decisions based upon that information, it is expected that educators and policymakers will be better able to answer such questions as:

- Are we getting our money's worth?
- How well is a new reform policy working?
- What are the impacts of policy changes on the system?
- Are we making adequate yearly progress toward our goals?
- How many students complete high school?
- How many graduates go on to college?
- How many graduates enter the workforce?
- Are there differences among ethnic groups?
- Are schools safe environments?

Ensure that equitable opportunities to learn are available to students. Another prevalent assumption about accountability systems is that they will encourage educators to create better opportunities for all students, including disadvantaged students. Indeed, the latest reauthorization of the federal Elementary and Secondary Education Act, which places a heavy emphasis on creating consistent statewide accountability systems, was entitled the No Child Left Behind Act. The assumption is that only by gathering data on students and disaggregating this information according to selected student characteristics can educators begin to diagnose and treat deficiencies in schools. And as noted earlier, researchers, including Scheurich et al. (2000), have argued that perhaps the most important impact of accountability on schools has been to focus the attention of educators and policymakers on whether schools are adequately serving at-risk students.

In McREL's work as a regional educational laboratory, providing services and resources to schools across the country, we often employ "logic models" — graphical depictions of "if … then" assumptions about our work (i.e., if we provide this service, then we expect this outcome). These logic models are designed to ensure that our efforts are focused on intended outcomes and that this is a logical causality between efforts and intended outcomes. Similarly, policymakers and educators may find it beneficial to make their purposes for accountability systems explicit in order to determine whether these systems have the necessary components or characteristics to

ensure their desired outcomes. In the final section of this report, we provide one logic model for accountability systems.

ALIGNING ACCOUNTABILITY SYSTEMS WITH PUBLIC GOALS FOR EDUCATION

In addition to being clear about the underlying purposes of accountability systems — *why* they were created — policymakers and educators also need to define the outcomes they are attempting to achieve — that is, *what* they are holding schools accountable for accomplishing. In most states, academic standards represent a general consensus of the desired academic outcomes, but surveys and conversations with members of the public reveal that non-academic outcomes are often equally important.

For example, in 1999, ICR Research Group found that the public's top three concerns relative to schools in their communities were lack of parent involvement (55%), use of alcohol or illegal drugs (51%), and undisciplined and disruptive students (50%). In citing these findings, SEDL researchers (Pan & Mutchler, 2000) noted that "these priorities contrast sharply with the predominant education reform agenda which, since the 1980s, has focused solidly on defining and measuring student and school performance" (p. 13).

In other words, while educators and lawmakers have spent the past decade focused on academic outcomes — how to set and measure them and what to do with the results — the public has had other, deeper misgivings about public schools. Parents want their schools to be accountable, but for more than just academic outcomes. Accountability systems need to be designed accordingly — to provide incentives for schools to attend to students' academic growth as well as their growth in other, nonacademic areas.

One key early finding from McREL's National Dialogue on Standards-based Education (Goodwin, Arens, Barley, & Williams, 2002) is that although parents and members of the public appear to generally support the idea of holding schools accountable, they may have a different view of accountability than policymakers. During focus group sessions, many parents expressed frustration with their schools' lack of responsiveness to their concerns or failure to treat them or their children as "customers." Many complained that their schools had become inaccessible "bureaucracies" or "government schools," not public schools. In short, when they talk about accountability, parents appear to be focused on making schools more responsive, not to policymakers or state officials, but, rather, to them. Thus, a system that makes schools more accountable to state officials, instead of parents and local community members, may run counter to what parents and members of the public are demanding.

Thus, it is important to consider *to whom* schools should be accountable. Given that half of schools' funding typically comes from state coffers and state legislators are responsible for the wise allocation of those dollars, it's reasonable to assume that schools should be, at least in part, accountable to state lawmakers and public officials. At the same time, "there is no escaping the logic which insists that public schools can't exist without their publics" (Mathews, 1996, p. 76).

Accordingly, accountability systems should be designed in a way that avoids a further widening of the rift that exists in many communities between the public and its schools.

Once the purposes of accountability systems have been examined and clearly specified, the next step is to identify ways of accomplishing those purposes. The following section synthesizes recommendations, guidelines, ratings systems for state accountability systems, and some research findings to identify the components and characteristics of effective accountability systems — that is, strategies that appear to be best able to accomplish the stated purposes of accountability systems.

CHARACTERISTICS OF EFFECTIVE ACCOUNTABILITY SYSTEMS

Numerous recommendations have been made regarding how accountability systems should be constructed, (e.g., Sirotnik & Kimball, 1999; Goff, 2000; Baker, Linn, Herman, & Koretz, 2002; Walberg, 2002). In addition, two organizations, Education Week and The Princeton Review, now annually review state accountability systems. An analysis of and synthesis of this mixed collection of recommendations, guidelines, ratings systems, and research (see Appendix A) reveals that although there are some significant differences among them, some consensus nonetheless emerges about the components or characteristics that accountability systems should reflect. Through this literature review and synthesis, we identified the following 12 essential components and characteristics of accountability systems (listed generally in order of frequency cited):

- Clear standards and expectations
- High-quality assessments aligned with standards
- Multiple measures
- High expectations for all students
- Readily understandable to the public
- Diagnostic uses for data
- Sanctions and rewards linked to results
- Flexibility and fairness to allow for local differences and creativity
- Alignment of resources, support, and assistance for improvement
- Balanced, comprehensive design
- Stakeholder support/engagement
- Fairness provisions

The following sections discuss each of the characteristics in more detail, and where possible, offer an analysis of the extent to which research findings support, refute, or are silent on their importance in improving student learning. It should be noted that in light of the fact that most state accountability systems are relatively new, there is not yet an extensive research base on their efficacy.

Researchers examining accountability systems are further hampered by the fact that the key intended outcome of "new accountability" systems is improved student learning, but real

student learning may not necessarily be synonymous with higher test scores. Indeed, many have wondered if gains on statewide assessments are less a reflection of students becoming better learners and more a reflection of students becoming better test takers. During the 1980s, for example, most states reported gains in norm-referenced test scores. The commonly cited reasons for these widespread rise in test scores, dubbed the "Lake Wobegon" effect (Cannel, 1987; Linn, 2001), included the reuse of the same tests year after year, use of old norms, and higher exclusion rates than in the norming samples (Linn, 2001). Anne Lewis (2001) describes the Lake Wobegon effect as follows:

> … what researchers have been trying to point out to anyone who would listen [is] the longer students take the same tests, they better they do. Texas, California, Florida and other states may certainly claim that the test scores show that their students are learning more. However, introduce a new test in any of those states, and Sisyphus' rock crashes to the bottom of the hill. And the climb must start all over again. (p. 567)

Thus, gains on large-scale exams may not translate into real gains in student learning. Researchers often note the importance of conducting studies to evaluate the degree to which gains in proficiency might generalize to other assessments. For example, although the percentage of Kentucky fourth graders scoring at the "apprentice level" or higher on the Kentucky Instructional Reporting and Information System (KIRIS) increased from 68 percent to 98 percent between 1993 and 1998, the percentage of Kentucky students scoring at the "basic" level on a different test, the National Assessment of Educational Progress (NAEP) actually *decreased* during the same period, from 58 percent to 56 percent (Linn, 2001). Similarly, results on the Texas Assessment of Academic Progress (TAAS) showing a narrowing of the achievement gap between minority and non-minority students were not replicated on the NAEP. In fact, even though minority students' NAEP scores rose, the achievement gap actually *increased* slightly. Linn (2001) argues that this "divergence in trends does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizability of gains reported on a state's own assessment" (p. 29).

Others, such as Craig Jerald (2001), contend, however, that the fact that not all of the gains reported on a statewide tests, such as the TAAS, are reflected on the NAEP, does not necessarily mean that the TAAS gains are meaningless:

> Teachers align their instruction to the test meant to measure the state's own curriculum and for which they are being held accountable. … Far from "dumbing down" their instruction or narrowly "teaching to" TAAS, the state's [improving] NAEP results suggest that Texas students are making real progress in mastering the kinds of challenging content and skills measured by tests like NAEP. (pp. 18–19)

Nonetheless, state policymakers, educators, and education researchers need to be mindful of the fact that rising scores on a statewide assessment are not necessarily a guarantee of real

gains in student learning or improvements in instruction. States should carefully consider the content assessed by their tests and how that converges with or diverges from other assessments. Such a purposeful examination of content would provide evidence to support differential growth across measures and evidence to support the technical validity of the assessment. Moreover, when evaluating the effectiveness of accountability systems and various characteristics of these systems, it may be necessary to consider more than just student achievement scores.

CLEAR STANDARDS AND EXPECTATIONS

One of the most often cited essential components of accountability is clear standards and expectations. WestEd researchers, for example, contend that "the foundation for any accountability system is a set of clearly defined content standards that spell out what students should know and be able to do" (Guth et al., 1999, p. 15). Many writers, including Walberg (2002) and Reeves (2002), echo this sentiment, arguing that the point of accountability systems is to get everyone in the education system focused on the same thing — improved student learning. In keeping with this notion, it only stands to reason that states need clear, well-written standards to guide improvement efforts.

As a result, a key question asked by those rating and evaluating accountability systems is if standards are clear enough to provide sufficient curricular guidance to educators. Fully 25 percent of Education Week's (2002) ratings for statewide accountability systems, in fact, are based on the inclusion of standards that are clear and specific in four core subject areas. Similarly, 6 percent of the Princeton Review's (2002) national ratings of state accountability systems are based on whether standards are "granular enough that a small number of test items can reasonably measure a students' mastery of that granule" (p. 3).

RAND researchers (Grissmer & Flanagan, 1998) offer some evidence that clear standards are important to raising student achievement. They singled out two states that demonstrated the most dramatic gains in NAEP scores between 1990 and 1997 — Texas and North Carolina — and concluded that "the most plausible explanation for the test scores gains are found in the policy environment established in each state" (p. i). Grissmer and Flanagan identified eight policy practices common to both states, one of which was "clear teaching objectives through statewide standards" (p. iii). In addition, they found that in both states "efforts were made to align the professional development, textbooks and curriculum with the statewide standards" (p. iii).

HIGH-QUALITY ASSESSMENTS ALIGNED WITH STANDARDS

Once clear academic standards have been identified and made public, authors of most accountability recommendations and guidelines contend, student progress on those standards needs to be measured using assessments aligned to those standards (see e.g., Guth et al., 1999). Most of the guidelines also heavily emphasize the quality of these examinations. Generally speaking, assessment quality involves three underlying issues: (1) reliability, (2) validity, and (3) alignment.

Reliability. Reliability refers to the ability of the test to accurately and repeatedly measure students' true ability in a specific content area. Though this may seem straightforward, there are many complicating factors. For starters, since it would require a tremendous amount of time for a student to be tested on all the subject matter in a given content area, the items on a test are only a sample of the possible items. The sampling process naturally leads to some degree of variability in the accuracy or reliability of the test. Also, students' performance varies based on how well they feel on the day of the test, how much effort they put into the test, and their level of alertness (Traub, 1994; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). These factors all contribute to a degree of error in the final score. Because all assessments have some degree of measurement error, test developers must acknowledge it so that users may make an informed decision about whether the level of error is acceptable.

Linn (2001) points out the common assumption that tests are more accurate in measuring student knowledge than they really are. For example, the Stanford Achievement Test, a test widely regarding as having a high level of reliability, nonetheless has a 30-point margin for error. Though most people assume that a score of 530 is better than a 500, the two scores are "well within the bounds of what would be expected as the result of measurement error" (p. 19). But on other tests, with far poorer reliability, the margin for error can be much greater. For example, on some standardized tests, a student who is *below* average could score as high as the 70^{th} percentile (Linn, 2001).

Such large margins for error, of course, create concerns for accountability systems. For starters, students can easily be misclassified if they score slightly (or not so slightly, depending on the tests' degree of reliability) above or below the cut-off point for a basic, proficient, or advanced scores on the statewide test.

Measurement error is not only present in student scores, but is also a concern with building-level achievement data and therefore should be publicly reported. This error is in part due to the sample of students who took the test as well as the items on the assessment (Linn, 2001). Measurement error becomes more of an issue for small schools whose scores can be greatly influenced by several high- or low-performing students. These high and low performers can not only skew the results for a given year, but also skew the results used to calculate for adequate yearly progress (AYP). For example, if a school has a small fifth-grade class of 15 students, but the overall school population is mobile, the school classification might be dramatically influenced if several very high- or very low-achieving students leave the school. Therefore, when calculating AYP, consideration must be given to the sample size of students in the school in order to generate a reliable statistic.

CRESST researchers (Baker et al., 2002) have argued that in their standards for accountability systems states should

- develop "rules for determining adequate progress of schools" that "avoid erroneous judgments" and
- publish "error rates associated with misclassification of individuals or institutions" along with test results. (pp. 2–3)

Validity refers to the degree to which a test measures the desired underlying construct and the appropriateness of the inferences made from the test scores (Messick, 1989). This definition clarifies not only the need for stakeholders to understand the content of their test, but clarifies the importance of developing guidelines regarding appropriate uses of assessment. Obviously, since no test can cover all the knowledge and skills that students are expected to learn in a given year, a sampling of that content is covered. Thus, a key question related to test validity is whether student performance on that small sample of questions adequately reflects their true knowledge of the broader curriculum. Moreover, a bigger question may be whether valid inferences can be made about the quality of instruction provided at those students' schools based on their performance on these test items.

Popham (1999), in fact, argues that standardized student assessments are *not* a valid measure of education quality, asserting that "employing standardized achievement tests to ascertain educational quality is like measuring temperature with a tablespoon" (p. 10). He also contends that a key shortcoming of using standardized tests — like those many states use for their statewide assessments — is "psychometric tendency to eliminate important test items." Standardized tests are designed to produce "comparisons among students from only a small collection of items." (p. 10) Popham explains:

> A test item that does the best job in spreading out students' total-test scores is a test item that's answered correctly by about half the students. Items that are answered correctly by 40 to 60 percent of the students do a solid job in spreading out the total scores of test-takers. … Test items answered correctly by 80 percent or more of the test takers, therefore, usually don't make it past the final cut when a standardized achievement test is first developed, and such items will most likely be jettisoned when the test is revised. (p. 11)

Yet such test items might cover important knowledge. In fact, teachers are more likely to devote instruction time to areas of the curriculum regarded as most important, and students, in turn, are more likely to answer related questions correctly. In order to construct standardized tests that effectively spread out student test scores, however, questions covering these areas are more likely to be discarded.

This tendency in standardized test construction gives rise to concerns that teachers will "teach to the test" by exclusively focusing on this smaller (and, most likely, idiosyncratic) sample of content on the test. As CRESST researchers (Baker et al., 2002) note, "learning the answers to the items on a single form by focusing exclusively on those items is not the same as learning the material for the domain of content the test is intended to measure" (p. 3).

Accordingly, to ensure the validity of inferences made from a standardized test, they recommend that "multiple test forms should be used when there are repeated administrations of an assessment" (p. 3).

Moreover, Popham (1999) argues that student assessments are really designed to measure how much students have learned — not necessarily the quality of the educational environment. He notes three factors that contribute to students' scores on standardized achievement tests: "(1) what's taught in school, (2) a student's native intellectual ability, and (3) a student's out-of-school learning" (p. 11). According to Popham:

> These sorts of items, because they tap innate intellectual skills that are not readily modifiable in school, do a wonderful job in spreading out test-takers' scores. The quest for score variance, coupled with the limitation of having few items to use in assessing students, makes such items appealing to those who construct standardized achievement tests. (p. 11)

Popham contends that "items that primarily measure differences in students' in-born intellectual abilities obviously do not contribute to valid inferences about 'how well children have been taught'" (p. 11). Clearly, using these items to judge educational effectiveness is inappropriate.

In light of such concerns, one of CRESST's standards for accountability system is that "if tests are to help improve system performance, there should be information provided to document that test results are modifiable by quality instruction and student effort" (Baker et al., 2002, p. 3). In other words, "tests need to be sensitive to differences in instructional quality and student effort" (Baker et al., p. 3), not simply socio-economic status, innate intelligence, or the acquisition of better test-taking techniques.

Some, though by no means all, researchers insist that the most valid assessments are ongoing classroom-based assessments. Sirotnik and Kimball (1999) for example, state:

> If you want students to be able to write, they have to write! And you have to assess it. It's much more expensive, but you can find out more (not all) of what you need to find out. Solving math problems is more complicated than merely answering multiple-choice items; you actually need to have students solve problems and explain what they are doing and why.
>
> These are not only better assessment practices; they are better teaching and learning practices (assuming you're a fan of higher-order thinking skills and the like). (pp. 212–213)

Alignment between standards and assessments is a third assessment-related issue raised by researchers and evaluators of accountability systems. The key issue here is whether a state's

tests actually mirror its standards. Both Education Week (2002) and the Princeton Review (2002) address this issue in their ratings of statewide accountability systems; whether states have criterion-referenced assessments aligned to state standards accounts for 12 percent and 10 percent, respectively, of their grades for accountability systems.

In their study of student achievement gains in North Carolina and Texas, RAND researchers found that having "state-wide assessments closely linked to learning standards" was a practice common to both states, noting that "new statewide assessment tests were … developed in both states which reflected the standards at each grade" (Grissmer & Flanagan, 1998, p. iv). Likewise, a meta-analysis of school factors associated with gains in student achievement conducted at McREL (see Marzano, 2000) found that "providing a well-articulated curriculum aligned with assessments" was the factor with the highest correlation to student achievement. According to McREL's analysis, when curriculum is (a) well-articulated, (b) aligned to assessments, and (c) school leaders monitor the extent to which it is actually covered, the measurable impact — or effect size — of such strategies is *31 percentile points* in student achievement.

Although alignment to standards is a critical component of an effective assessment, it's important to recognize that not all academic standards are easily measurable. States should be explicit about which standards are measured and the weight each standard is given on the assessment. For example, the state of Colorado has a reading standard that requires students to demonstrate knowledge and skills in reading, writing, speaking, listening, and viewing. Though Colorado certainly assesses many of these skills, speaking skills are not measured on the state assessment because of the format of the test.

MULTIPLE MEASURES

Given the concerns about test reliability and validity, many guidelines and models of accountability systems call for the use of multiple measures (Sirotnik & Kimball, 1999; Baker et al., 2002; Education Week, 2002; The Princeton Review, 2002). Sirtonik and Kimball, for example, argue that an "accountability system must not be driven by a single indicator (e.g., test scores) and simplistic formulas for rewards or sanctions based on that indicator" (p. 211). Sirtonik and Kimball offer the following justification for demanding the use of multiple measures:

> Common sense suggests that scores on one test (which itself is only a sample of many possible performances) cannot possibly represent all that is going on in a school, any more than the temperature reading on a thermometer can represent all that is going on in a human body. Making inferential leaps from test results to the causes of those test results is foolish in the extreme. What other data — even just a minimal set — might be useful in understanding the average test score posted by a given school? (p. 211)

Similarly, CRESST researchers (Baker et al., 2002) maintain that for several reasons, a single test should not be the basis of decisions about individual students:

> First, no test is perfectly reliable. There is always a degree of uncertainty associated with any test score. That uncertainty needs to be taken into account when making decisions about individual students. Second, all tests have less than perfect validity. Hence, it is important to look for other information that will either support or disconfirm the information provided by a single test score. The importance of obtaining other information to confirm or disconfirm the information provided by a single test score increases as the importance of the decision and the stakes associated with it increases. Yet another reason for multiple sources of information is the limitation of a single measure as a sample of the domain(s) of interest. (p. 3)

Therefore, CRESST researchers assert, other kinds of data about students should be considered, including information related to attendance, mobility, retention, dropouts, and graduation. Similarly useful is information about instructional resources and students' opportunities to learn the content specified in academic standards and curriculum materials.

HIGH EXPECTATIONS FOR ALL STUDENTS

In keeping with the notion that accountability systems were created, in part, to ensure greater equity among students' opportunities to learn, a key component of most accountability systems is high expectations for all students. As Baker et al. (2002) assert, "accountability systems should include the performance of all students, including subgroups that historically have been difficult to assess" (p. 2). The basis of the argument for testing all students is that only by including all students in the assessment system can states ensure that schools and teachers focus their efforts on helping every child learn. RAND researchers (Grissmer & Flanagan, 1998), in fact, found that Texas and North Carolina both place great emphasis on holding all students to the same standards, noting, in particular, that disadvantaged students and advantaged students are held to the same standards.

READILY UNDERSTANDABLE TO THE PUBLIC

Many argue that in order to adequately inform students, parents, and teachers about student progress, accountability systems must produce data and results that are meaningful to students, parents, and teachers. For example, the Princeton Review's (2002) evaluation of state accountability systems reflects this notion: "Is performance data shared with the public along with explanation and contextual detail appropriate for a general audience?" (p. 4). Similarly, Walberg maintains that a key principle of accountability systems should be "user-friendliness." "Readily understood reporting is desirable," he writes. "Perhaps even a single number or two may best serve occasionally. … What isn't as useful is a mass of undigested numbers often reported by states and districts in large, unwieldy books of computer printouts" (p. 158).

Although it may be helpful to members of the public to combine data into a single or small set of numbers, CRESST researchers (Baker et al., 2002) argue that when such an approach is taken, states should make explicit how different types of test content and other elements in the system are weighted. They also insist that "results should be made broadly available to the press, with sufficient time for reasonable analysis and with clear explanations of legitimate and potential illegitimate interpretations of results" (p. 5). In support of this standard for accountability systems, Baker et al. offer the following observation:

> The press plays an important role in the interpretation of the results produced by accountability systems. Legitimate interpretations of results require an understanding of what goes into them and some of their technical characteristics. Those responsible for the accountability system also have a responsibility to help ensure proper interpretation of the results and to minimize inappropriate interpretations to the extent possible. Efforts to assist the press in understanding the results, their strengths and limitations, and the legitimate and illegitimate interpretations can pay considerable dividends in improved coverage by the press and better understanding by the public. (p. 5)

DIAGNOSTIC APPLICATIONS

Another key purpose of accountability systems is to provide schools with information they can use to guide their decision making. Though test results can help focus educators on what's important — student learning — in some places, there has been a tendency to focus only on test scores and ignore other data that schools could use to guide improvement efforts. Throwing the baby out with the bath water, so to speak, and only reporting test scores, will leave schools in a position of knowing they need to improve, but not knowing how or what needs to change. Indeed, Fuhrman (1999) found this to be true of many schools labeled as inadequate. The label created a state of urgency in the school, but educators had little guidance about where to channel their energies. To avoid creating this sort of unguided urgency, it's important to consider all the data that states, districts, and school might collect and the purposes this information might serve — and to help those agencies understand their data and learn how to make decisions based on their data.

In order to provide schools and teachers with diagnostic information, data must be specific enough to be of use to educators as they identify and select corrective actions. CRESST's standards, for example, call for accountability systems to "include data elements that allow for interpretations of student, institution, and administrative performance" (Baker et al., 2002, p. 2). The key, according to Reeves (2002), is providing school communities with sufficient data to be able to determine what works:

> If I go to one more conference where we hold hands and chant, "All children can learn," I'm not going to be able to take it anymore. I believe that all children can learn, but I have never achieved anything with a mantra. Accountability is not about chanting mantras; it's not about generalities.

> We've got to know specifically what works. We've got to investigate which strategies in our own communities are specifically associated with improved student achievement. (p. 1)

Educators also need to receive the outcome data promptly enough to make needed changes. Walberg (2002), for example, notes that "school boards and teachers often get test results long after their time of prime usefulness, namely immediately" (p. 158). Just as "one mark of a good teacher is getting the test back the next day," he argues, state accountability systems should strive to do the same (p. 158).

Reeves (2002), similarly, argues for timeliness, citing the "Nintendo Effect," an analogy coined by Jeff Howard, president of the Efficacy Institute. Howard notes that even the most easily distracted child will sit for hours in front of a Nintendo game. He asks, "How long would that child be staring at the screen if you mailed his Nintendo scores to him nine weeks hence?" In short, part of what keeps children glued to the game is its immediate feedback. Reeves argues:

> If we're going to build a holistic accountability system, once-a-year feedback is not sufficient. We should be building a system that every month gives feedback to our children, our leaders, and our teachers so we can get busy building better instructional systems." (p. 1)

Walberg (2002) believes those days may not be far off, citing as evidence the fact that "more than 100 firms are now working on computerized tests administered on the Internet" (p. 159). He believes that Internet tests could be, and perhaps should be, the heart of future accountability systems:

> Such tests can be scored in several seconds; they can save on printing costs, can be quickly updated, and may require as few as a third of the testing time and items as the usual tests because they adapt the difficulty of the items to the students' ability, which is better estimated with each successive item. Open to the public, parents and students, independent of schools, could check their progress on demand in any given subject. (p. 159)

RAND researchers, in fact, identified "computerized feedback systems" — albeit less sophisticated ones — as reform elements also common to both Texas and North Carolina. They noted that "both states have a well-designed computerized system of storing the testing information and providing access to it in various ways to teachers, principals and school districts" (Grissmer & Flanagan, 1998, pp. iv–v).

LINKED TO SANCTIONS AND REWARDS

Using sanctions and rewards to spur school improvement efforts, though not as universally supported as some system elements, nonetheless surfaces in many accountability recommendations and guidelines. The argument in favor of sanctions and rewards is basically

that, as Walberg (2002) puts it, "simply publishing results appears insufficient for progress … Schools of choice risk closing if they attract no students. Analogous thinking dominates much of society. Why not schools?" (p. 159). Accordingly, Education Week's (2002) ratings of state accountability systems includes whether the state holds schools accountable for performance through sanctions, including closure, reconstitution, transfers, withhold funds, and rewards.

RAND researchers (Grissmer & Flanagan, 1998) found that both Texas and North Carolina have performance-based financial rewards as well as "the power to disenfranchise school districts and remove principals based on sustained levels of poor performance" (p. iv). They note, however:

> A key issue faced by states in establishing systems of accountability is how to take into account the strong correlation of test scores with the socio-economic status (SES) of the students. Perceived unfairness in the system of rankings and rewards can seriously erode the trust necessary for effective incentives. (p. 22)

In light of such concerns, both Texas and North Carolina take into account both "the absolute levels of test scores and the year to year gains in scores" (Grissmer & Flanagan, 1998, p. 23).

CRESST researchers (Baker et al., 2002) contend that sanctions and rewards should start out broad and diffuse, and then, as the system aligns, move to specific consequences for individuals and institutions:

> Starting with broad, diffuse stakes (e.g., public reporting of aggregate achievement results for schools) allows participants time to make the changes needed to meet expectations before being confronted with specific rewards or sanctions for performance (e.g., monetary rewards to schools or teachers, graduation requirements for students). Advance warning and phasing-in of stakes enhances both the perception of fairness and the actual fairness of the accountability system. (p. 5)

Not everyone agrees, however, with the need for sanctions in accountability systems. Sirtonik and Kimball (1999) argue against the punitive nature of accountability systems:

> Through injections of human and fiscal resources, the accountability system must nurture and support districts and schools in decline or those making little or no progress. Behavioral psychology has taken a bad rap lately, in light of developments in cognitive psychology and the popularization of constructivism. Yet we recall some pretty sound principles of learning theory that we suspect still hold true today. One is that punishment doesn't work very well in changing behavior. What works a whole lot better is counter-conditioning — finding ways to reinforce desired behaviors that interfere with less desired ones. (p. 213)

FLEXIBILITY

Much of the current rhetoric surrounding new, outcome-oriented accountability systems is that they should be designed with the following trade-off in mind: results for flexibility. That is, the system is said to demand results from schools and districts, yet leave it up to schools and districts how to achieve those results. For example, RAND researchers (Grissmer & Flanagan, 1998) found that another feature common to both Texas and North Carolina is increased local control and flexibility for administrators and teachers:

> In both states, many statutes governing schools and teaching were repealed. Fewer constraints were placed on district superintendents and principals on how money is spent. The clearly expressed policy was to allow schools locally to take different approaches to achieving the objectives. The state departments of education were downsized in both states, and refocused to the assessment and accountability programs — measurement and reporting scores and organizing resources for poor performing schools and school districts. In both states there is some evidence of "cottage industries" developing in the private sector, which provide various kinds of services and support for the school systems. These firms may provide supplies, training, curriculum, and computerized learning systems. The newer funding flexibility given to local school districts and schools may be partly responsible for the emergence of such firms. (p. 24)

Similarly, one of Sirotnik and Kimball's (1999) standards for standards-based accountability is that the system "must be based on high-quality content standards that allow districts, schools, and teachers to be creative, flexible, and thoughtful in constructing and delivering a curriculum that meets the standards but is not so narrow that it limits the rich array of curricular experiences and possibilities for teaching and learning in a multicultural and democratic society" (p. 212).

Fuhrman (1999) notes, however, that too often, accountability systems' focus on results have not been accompanied by flexibility:

> Bearing down on performance makes sense if the state simultaneously relaxes from process regulations that might restrict flexibility in reaching the new performance goals. Examples are mandates to teach students or teachers specific topics, like Lyme Disease prevention, that might be worthwhile but that are not related to the state's performance goals. Despite rhetoric about more autonomy in return for greater accountability, however, most states have not repealed much of the existing process regulation. (p. 5)

ALIGNMENT OF RESOURCES, SUPPORT, AND ASSISTANCE

Many models of accountability, including Education Week's ratings system, include the need for aligning resources and support with the goals of the system. RAND researchers

(Grissmer & Flanagan, 1998), for example, note that recent research makes the case that "resource levels can make significant differences in achievement," especially for disadvantaged students (p. 38).

In addition to equitable funding, researchers have argued that states must provide schools with support to help them develop the capacity they need to make improvements. Fuhrman (1999), for example, notes that research "shows that new accountability systems can be motivating," however, too often, "their design is based on one of two assumptions: either schools already possess the capacity, but not the will to meet goals, or, once the goals are clear and valued consequences are attached, schools will shop for or find the capacity they need to meet goals" (p. 9). She notes, however, that, "the mere imposition of a new accountability system… does not unleash some hidden capacity" (p. 9). In short, although accountability systems may create more *will* to succeed, they often fail to help schools find a *way* to succeed.

BALANCED, COMPREHENSIVE DESIGN

A common criticism leveled against many existing accountability systems is that they fail to take into account student motivation. That is, teachers and schools are often judged on tests that students have very little incentive to do well on. Thus, one of CRESST's standards for accountability systems is that the stakes "should apply to adults and students and should be coordinated to support system goals" (p. 4). They insist that failure to do so can have negative consequences for both teachers and students:

> If teachers and administrators are held accountable for student achievement but students are not, then there are likely to be concerns about the degree to which students put forth their best effort in taking the tests. Conversely, it may be unfair to hold students accountable for performance on a test without having some assurance that teachers and other adults are being held accountable for providing students with adequate opportunity to learn the material that is tested. (Baker et al., 2002, p. 4)

In other words, two key considerations in creating balanced, comprehensive accountability systems are *who* the system should hold accountable for *what* outcomes. Obviously, many people are responsible for the success of students — including state policymakers, district leaders, building leaders, community members, teachers, parents, and finally, students themselves. Many people have argued that a fair accountability system would seek to hold all of these players accountable, to some degree, for student learning. Such "reciprocal" accountability systems would recognize that student success is a product of several players, from teachers to parents and guardians to the entire community to the students themselves taking responsibility for their learning.

Former Ohio State Superintendent of Public Instruction John Goff (2000) has created a blueprint for "comprehensive accountability" systems. Goff's system defines a "responsibility schema" in order to hold the players responsible for educating children accountable for doing so.

His schema includes students, teachers, parents, principals, superintendents, local boards, community members, teacher preparation institutions, governors, legislators, chief state school officers, state boards, the business community, and the media.

To a large extent, ensuring that accountability systems are "reciprocal" or "comprehensive" is an ethical concern. As a simple matter of fairness, accountability systems should avoid making scapegoats out of any particular group. But it is also reasonable to assume that comprehensive or reciprocal accountability systems may be more effective because they apply pressure on more "levers" in the system.

STAKEHOLDER SUPPORT/ENGAGEMENT

Some researchers have argued that another vital component of accountability is stakeholder support and engagement. WestEd researchers (Guth et al., 1999), for example, argue for engaging stakeholders in designed and maintaining the accountability system:

> Since intervention programs may have significant sanctions associated with them, it is important that there be broad-based support for such an approach. In fact, involving affected stakeholders in all stages of developing an accountability system from standards development to carrying out of sanctions requires constructive communication with and buy-in from students, parents, teachers, school administrators, district policy makers, business, and other community members. (p. 16)

FAIRNESS PROVISIONS

In light of the fact that no assessment is perfectly reliable or valid, coupled with the consequences now tied to accountability systems, CRESST researchers say that states should make "appeal procedures available to contest rewards and sanctions" (p. 4). Similarly, the Princeton Review (2002) asks whether "students have the opportunity to retake the test, if necessary" and whether there are "due process guidelines for people accused of cheating" (p. 4).

Fairness in assessment is a complex and controversial topic that is the subject of much discussion. Multiple issues impact every aspect of test development, test implementation, and test use. For example, test developers must determine if items function differently or are worded in a manner that might discriminate against particular groups of students. Items that are biased do not measure differences in students' ability in the subject area but reflect differences in cultural or social experiences. Likewise, educators must ensure that testing processes treat students equally, for example by allowing appropriate accommodations, standardizing test administration procedures for students, and keeping individual test results confidential. Finally, and perhaps most important, a fair test only measures students on what they have had the opportunity to learn. This is perhaps the most difficult of issues to address, particularly at the state level. All schools and districts must ensure that their students are learning material that is tested by the state. Such

state oversight often overrides local control but also ensures that all students are engaged in instruction that meets state standards.

FRAMEWORKS FOR COMPREHENSIVE ACCOUNTABILITY

This section provides three models for conceptualizing the various components of accountability systems and the relationships of these components. By extension, these conceptual models could also serve as the basis for constructing evaluations of standards-based accountability systems. These models, especially the latter two, offer a broader view of accountability than what we found in most of the recommendations, guidelines, and models we reviewed. Factors such as "orderly atmosphere," "time on task," "readiness for school," and "societal support for learning" are purposefully included in our frameworks even though they rarely appear as factors in existing accountability systems. We believe that collecting and reporting data on these factors can offer a richer picture of the health of the system. Most likely, data on these factors may already be collected in state and regional accreditation systems, teacher evaluations, program evaluations, and other mechanisms. Thus, it may not require much additional effort to incorporate these data into richer, more comprehensive accountability systems.

A PURPOSE-ORIENTED ACCOUNTABILITY SYSTEM

The first model, provided as Figure 1, is a logic model designed to depict the relationships between the key, underlying purposes of the accountability system and its components or features. Such a model could help policymakers determine the extent to which their accountability system has been designed to accomplish its key purposes. Viewing the graphic from the top down, it shows that the overarching purpose of the system is to improve student learning. Five secondary purposes of the system are depicted in larger arrows directly below the overarching outcomes. The smaller arrows each depict different system features, which align with the secondary purposes of the system. The system is based on clear standards and expectations, which flow directly into quality assessments and multiple measures.

AN INPUT-PROCESS-OUTPUT ACCOUNTABILITY SYSTEM

The second model, shown in Figure 2, focuses more on the diagnostic purposes of accountability systems. That is, rather than viewing accountability narrowly as an elaborate system of incentives designed to compel or entice schools into improving, it focuses more on creating a sophisticated indicator system to help teachers, as well as school, district, and state leaders, identify specific shortcomings in their education systems. Although it still measures outcomes — namely, student achievement — it also collects data on various inputs, including teacher experience, per-pupil expenditure, and parental support. At the same time, it measures process variables, including school-level factors (e.g., the extent to which the school provides an orderly atmosphere, a quality school curricula, and exemplifies achievement-oriented policy). Such a system, of course, is more similar to an accreditation process than new accountability systems. However, given that accreditation reviews can yield valuable information that states can

use to help schools improve, we suggest that they can still play a vital and compatible role in current accountability systems, especially when the two systems are aligned.

AN ISSUE-ORIENTED ACCOUNTABILITY SYSTEM

The third model, shown in Figures 3a and 3b, is predicated on the identification of a broad array of issues related to successful public schools. Like the previous framework, this one also is focused on diagnostic assessment of the entire system. The comprehensive nature of this framework is evident in Figure 3a. The system depicted here seeks to provide data on not only the quality of student learning, but also a broader array of inputs that we know are crucial to student success — including school readiness and community and family support for students. Figure 3b focuses on one of these issues areas — learner outcomes — in order to provide more detail about the kinds of indicators that might be tracked within this sub-issue. This framework obviously differs from most current systems in that it focuses on much more than student performance on large-scale assessments; it attempts to assess the functioning and health of the entire system.

Figure 1. A Purpose-Oriented Accountability System

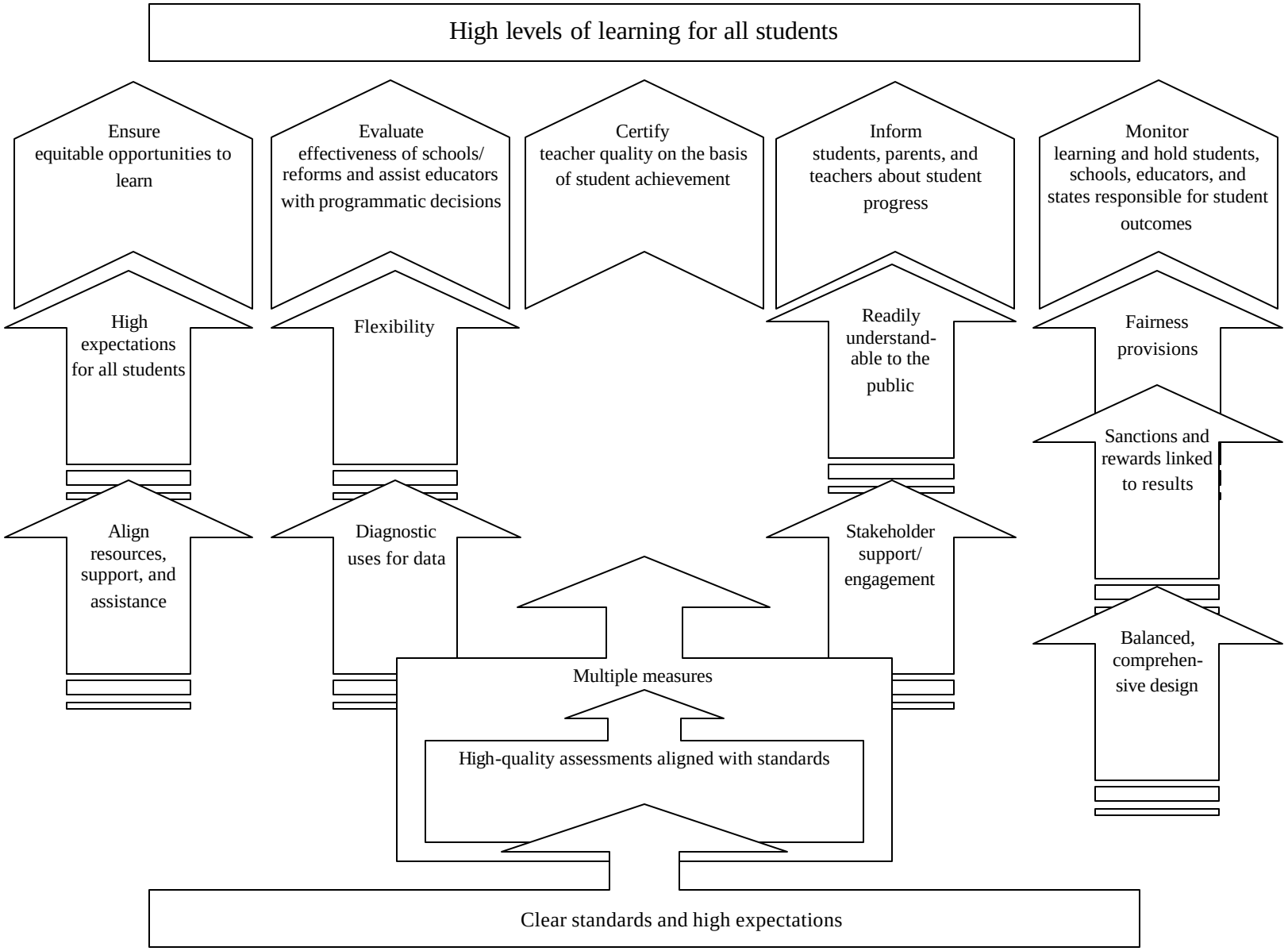


Figure 2. An Input-Process-Output Accountability System

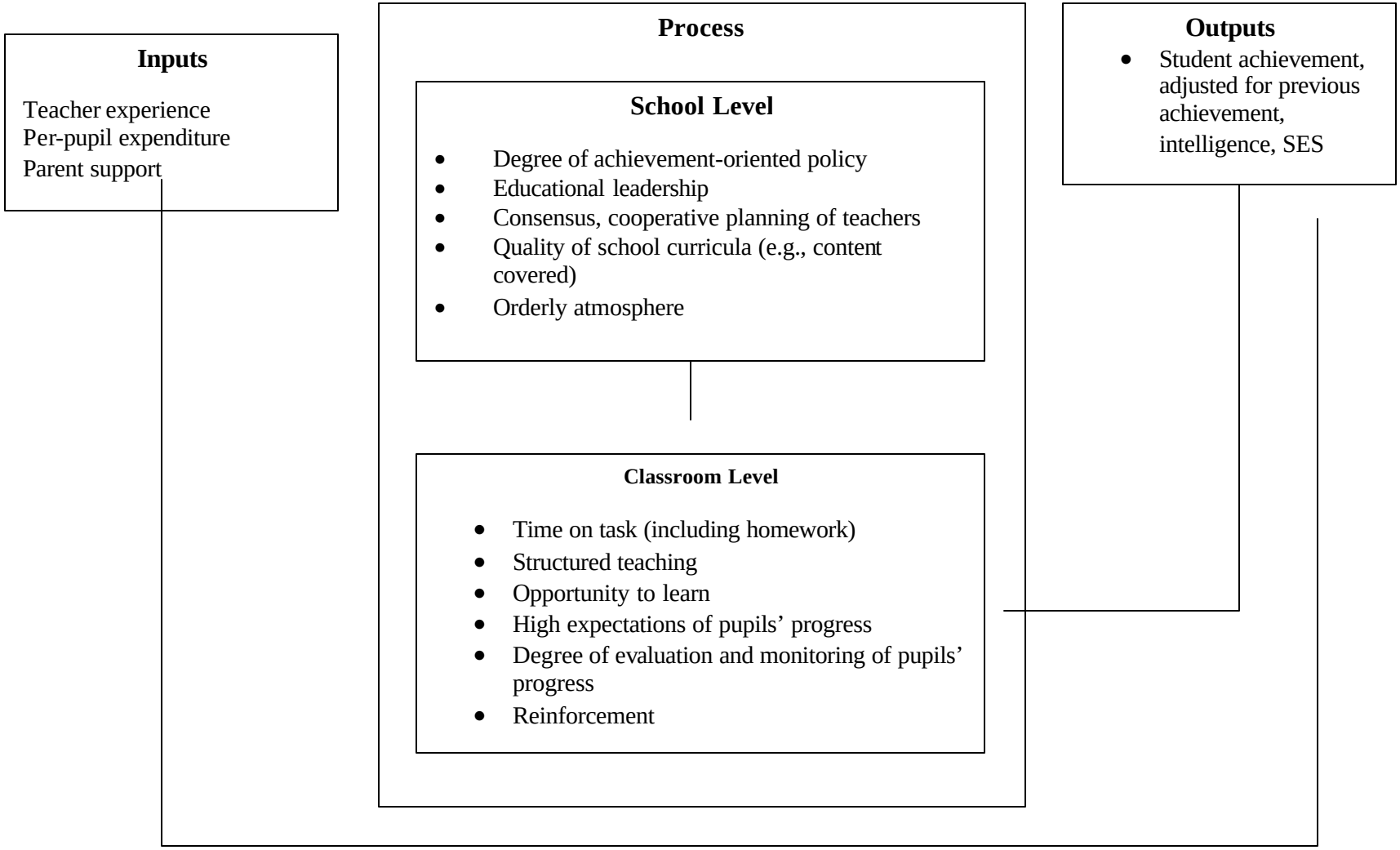


Figure 3a. An Issue-Oriented Accountability System

Primary Issue Level

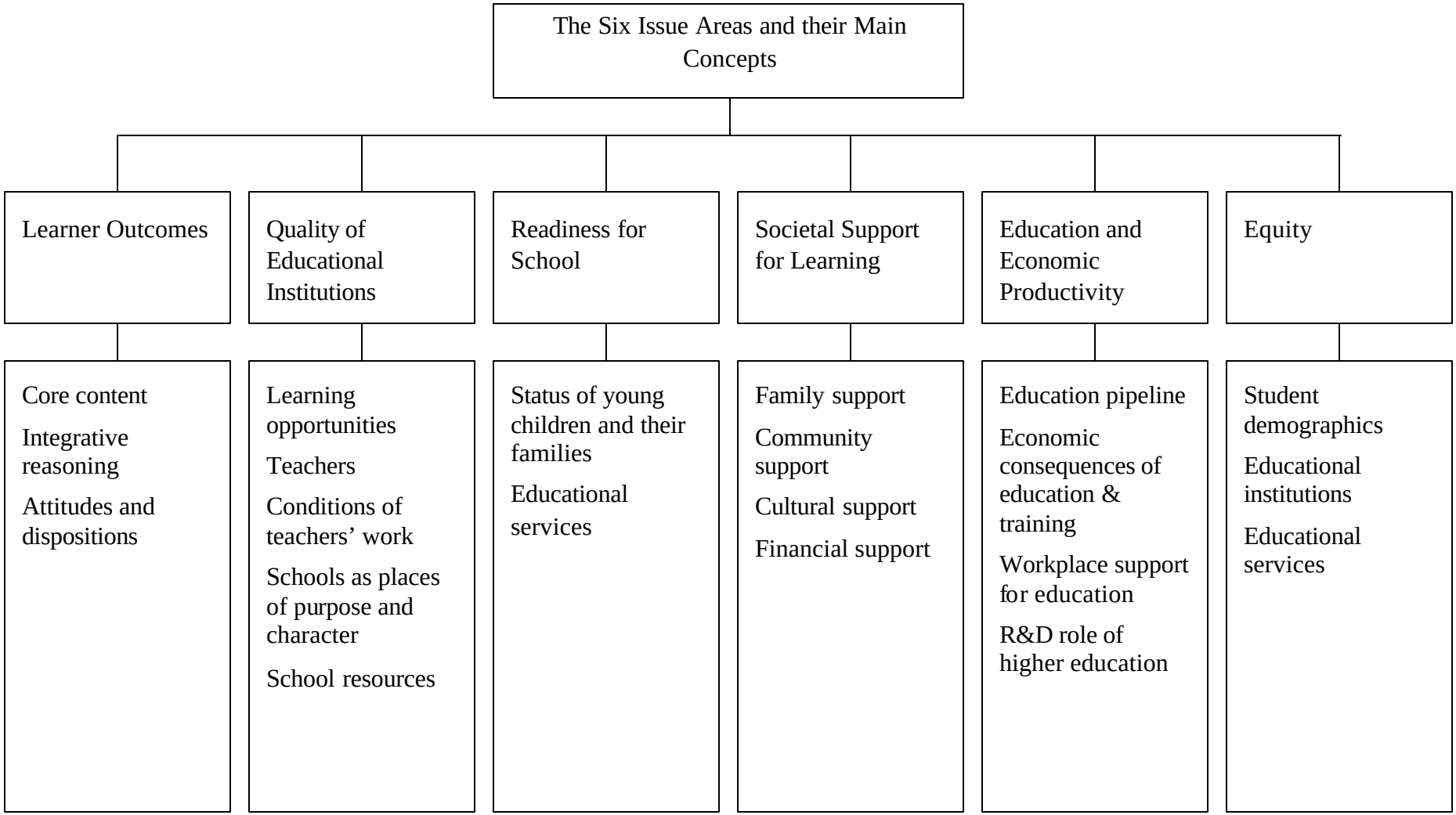
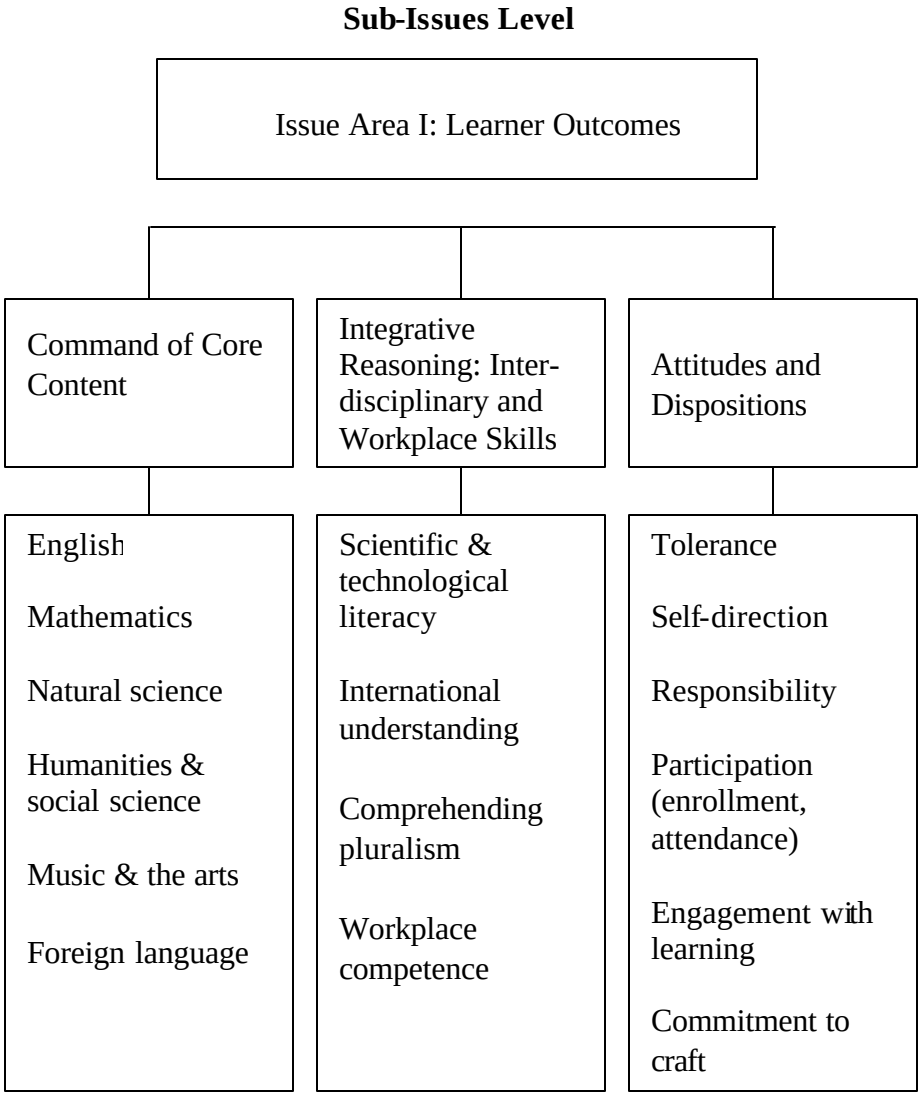


Figure 3b. An Issue-Oriented Accountability System



FINAL THOUGHTS

The previous frameworks, which offer a broader view of accountability than represented by most current models or recommendations for these systems, are derived from a program evaluation perspective. That is, they are conceived generally as evaluations of the systems' health, guided by the question, What information do policymakers, educators, parents, and the public need in order to know how their schools are doing? We believe this question should be the first one that policymakers ask themselves when designing accountability systems. Only after gaining a clear, comprehensive perspective on where the system is succeeding and where it is failing can we correctly determine *whom* to hold accountable for improving *which* aspects of the system.

It's worth noting that although the 50 states, with their separate governments and schools systems, are generally considered laboratories for democracy and policy experiments, during our review of existing criteria and recommendations for effective accountability systems, it became readily apparent that by and large, most states now have accountability systems that are remarkably similar. Differences exist, of course, but they are generally nuances — differing levels of specificity of standards and benchmarks, different types of test items and questions on statewide assessments, different kinds of sanctions and rewards applied to schools and districts. For the most part, though, all of the states' accountability systems are comprised of the same core components: statewide curricular standards, statewide annual assessments, and sanctions tied to performance on those tests (which generally is regarded as accountability).

In short, what is most remarkable about these systems is their striking similarity. Given that all of these systems were built from the ground up, one might expect that many radically different approaches to school accountability would have been developed. Instead, states appear to have followed a more cautious, bandwagon approach to designing their accountability systems. And now, in response to the No Child Left Behind Act of 2001, these systems may become even more standardized, with very few departures from the norm. This may be unfortunate, as it could deprive the nation of laboratories of the different forms of accountability that could be developed, tested, and then, if proven effective, scaled-up to other locations.

QUESTIONS TO CONSIDER

As a result of this uniformity of approaches to accountability systems, some of the following big questions regarding the design of effective accountability systems may go not only unanswered, but also unasked:

- Do content standards constitute all of the outcomes expected for students? Or are there equally important, but less easily measurable, outcomes that schools should focus on as well?
- Are standardized tests the only way for students to demonstrate their progress toward meeting standards?

- What other ways can accountability be created besides sanctions tied to student performance on standardized tests?
- How can school accountability be balanced with student and parent responsibility?
- What are the best incentives to encourage real and sustained school improvement?
- Are there effective ways of measuring the extent to which schools, communities, and taxpayers are providing all children with opportunities to learn?

TAKING A FRESH LOOK AT THE ELEMENTS OF ACCOUNTABILITY SYSTEMS

Given the paucity of alternative forms of accountability, it might be worth defining the essential elements of accountability systems in a broad way, and then creating new models based on the purposes of accountability identified earlier in this report. For example, one might argue that at the heart of all standards-based accountability systems are four main elements: (1) making goals and expectations explicit and public; (2) measuring individual and school progress toward those goals; (3) reporting individual progress to parents and school progress to the public; and (4) creating an incentive structure that encourages school improvement. One could then build a system by thinking more broadly about what each of these four elements might entail.

Making Goals and Expectations Explicit and Public

The first element, making goals and expectations explicit and public, often is interpreted narrowly, as academic standards. However, it also can be conceived more broadly. For example, the Chugach School District in Alaska has identified goals for its students in several areas, including academics, but also service learning, character development, and cultural awareness. Students are required to demonstrate progress toward meeting defined standards in each of these areas before graduating. In other words, all areas — even those that defy easily quantifiable measurement — are deemed important and necessary. Under many current accountability systems, however, only those learning goals that can be measured on standardized tests receive much attention from educators and school leaders.

Measuring Individual and School Progress Toward Defined Goals

Similarly, the second element of accountability systems — measuring student success — is also defined narrowly as achievement on standardized instruments, namely pencil-and-paper assessments. Some states, such as Vermont and Kansas, have experimented with portfolio assessments. Many people argue that although these tests often are more expensive and more prone to subjective interpretations, they represent a more authentic approach to assessment, one that encourages not only demonstration of higher-order thinking skills, but also greater student motivation and responsibility for their own learning. However, most states have opted for far less expensive standardized assessments, which come with the assurance of reliability and validity. The selection of such tests is often predicated on the notion that the results of these assessments

will be used to compare districts and schools in order to determine sanctions and rewards. Therefore, the test results must provide a common yardstick for all schools across the state.

Reporting Individual Progress to Parents and School Progress to this Public

The third element of accountability systems — reporting outcomes to parents and the public — also is often defined narrowly. For example, data reported by states and districts to the public are usually more oriented toward achievement data, rather than indicators of opportunities to learn. Individual student achievement data typically are not provided to parents or guardians; when this information is provided, it often is provided several months after the test administration, or it does not provide parents and guardians with an easily discernible snapshot of children's strengths and weaknesses. Moreover, these data are rarely aligned with student report cards sent from the schools, career and academic guidance from counselors, input during teacher conferences, and so on. In short, the information reported is of little diagnostic use to parents, educators, or policymakers.

Creating an Incentive Structure that Encourages School Improvement

Incentives have seemingly become the "name of the game" as far as accountability is concerned. More time and energy appear to have been devoted to identifying the right "levers" or incentives to produce change, rather than creating a careful, thorough, and fair evaluation of the system. Indeed, the notion of accountability has evolved to become inextricably linked with the concept of incentives, or, more precisely, sanctions or rewards.

It's also worth pointing out that like the other elements of the system, incentives also have become narrowly defined as sanctions or rewards tied largely to student test results. Some researchers, such as Sirotnik and Kimball (1999), argue against making accountability systems punitive because, they assert, positive consequences are likely to be more effective in changing behavior than negative ones. But even if everyone were to agree that rewards are more effective than punishment, it's not yet clear what kinds of positive rewards are most effective. For example, merit pay programs have typically received mixed reviews from researchers. Oftentimes, the rewards given to teachers in successful schools amount to less than $500, arguably not commensurate with the level of effort required to receive the bonuses. Or if they are designed to reward individual teachers, they wind up pitting teachers against one another for a limited pool of funds. As Kelley and Odden (1995) note, "Such competition among teachers works against the collaborative culture found in most highly effective schools and thus, is at odds with strategies to improve school performance" (p. 1).

All too often, we have noticed that those who argue against using accountability systems to shame or blame schools into improvement fail to provide an alternative mechanism for encouraging and, indeed, demanding school improvement. Accountability systems, were, in fact, borne out of a general sentiment that schools needed to improve and provide opportunities for all students to learn. And as Walberg (2002) and others have argued, simply reporting data publicly does not by itself provide educators and school leaders with ample incentives to improve. Indeed,

it appears that in spite of ample evidence that a great number of students were not being served well by our schools, school leaders, educators, and the public were content to maintain low expectations for those students and let them fall through the cracks. Thus, we would argue that the question is not *whether* accountability systems should include incentives, but rather what *mix* of incentives is most effective in encourage real, lasting, and systemic improvement? And who should provide these incentives? Should they all come from the state? What incentives should local boards provide to encourage school improvement?

Another way to think of incentives might be to ask the question, Whom do we trust to act upon the available information and foster changes to the system? Finn (2002) asserts that there are four answers to this question:

- (1) "trust the system" (i.e., create rules and procedures to guide inputs)
- (2) "trust the experts" (i.e., use professional norms and expertise to guide necessary changes)
- (3) "trust, but verify" (i.e., create standards, verify whether they are being met, and impose consequences to influence changes)
- "trust the customers" (i.e., let marketplace forces and parental choice drive changes to the system). (pp. 24–26)

Finn asserts that these four responses are not mutually exclusive — that is, accountability can be created through combinations of these strategies. The question, then becomes, Which of these strategies (or combinations thereof) is most likely to create desired changes in the system?

It may also be worth asking, What can federal, state, and local policymakers reasonably hold schools accountable for accomplishing? Do they need to provide both opportunities to learn *and* ensure that all their students take advantage of those opportunities? Or to borrow an old expression, should we require our schools to both lead the horses to water… *and* to make them drink? This issue, has, in fact, surfaced frequently during McREL's Nationwide Dialogue on Standards-based Education. Most participants in the May 2002 focus group sessions held in Kansas City, Missouri agreed that student success is ultimately the result of student motivation, personal responsibility, and parental/guardian support (Goodwin et al., 2002). Many believe that the "big question" policymakers need to consider is how to construct a system of incentives that provides a fair balance (one that avoids "blaming the victims") between holding schools and teachers accountable, while recognizing that ultimately students and their guardians are responsible for their learning.

One way to strike this balance might be to attach some incentives to schools' ability to demonstrate they are providing all students with capable teachers, safe environments, rigorous curricula, and so on — but not require student success. The rest would be left up to parents, students, human service providers, and communities. Or the real accountability mechanism might come largely through parental choice. For example, one might conceive of a system in which states and districts are required to ensure equitable per-pupil funding and measure and

report the success of students in each school. But accountability — that is, incentives to change — might come largely in the form of competition for students. Under such a system, it's conceivable that schools would become more accountable to parents and students (and, by proxy, taxpayers) than to policymakers.

NEXT STEPS

In 2003, we propose to examine ways to integrate these "big picture" questions into our framework for designing and evaluating accountability systems. In so doing, we hope to broaden the current conversation around accountability in order to encourage policymakers and educators to think creatively about their accountability systems and perhaps envision new possibilities for measuring and holding schools accountable for ongoing improvement.

In summary, this evaluation framework and process will attempt to tie together the various issues raised in this report by helping policymakers to

- re-examine their purposes for accountability systems;
- create a theoretical model for how the accountability system will accomplish its intended outcomes;
- use research to guide the inclusion of particular strategies in the accountability system; and
- ask "big picture" questions about the design of their systems and whether alternative, "outside the box" approaches might better accomplish the intended purposes of the system.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational psychological testing*. Washington, DC: American Educational Research Association.

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). Standards for educational accountability. *CRESST Policy Brief, 5*.

Cannel, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends of Education.

Cuban, L. (1990). Reforming again, again, and again. *Educational Researcher, 19*(1), 10–21.

Education Week. (2002, January 10). Quality Counts 2002: Building blocks for success [special report]. *Education Week, 21*.

Finn, C. E. (2002). Real accountability in K–12 education: The marriage of Ted and Alice. In H. J. Walberg (Ed.), *School accountability.* Stanford, CA: Hoover Institution Press, Stanford University Press.

Fuhrman, S. H. (1999). The new accountability. *CPRE Policy Briefs, RB-27*.

Goff, J. M. (2000). *A more comprehensive accountability model*. Washington, DC: Council for Basic Education.

Goodwin, B., Arens, S. A., Barley, Z. A., & Williams, J. (2002). *Understanding No Child Left Behind: A report on the No Child Left Behind Act of 2001 & its implications for schools, communities, & public support for education*. Dayton, OH: The Kettering Foundation.

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

Guth, G. J. A., Holtzman, D. J., Schneider, S. A., Carlos, L., Smith, J. R., Hayward, G. C., et al. (1999). *Evaluation of California's standards-based accountability system*. Menlo Park, CA: WestEd, Management Analysis and Planning, Inc.

Jerald, C. J. (2001). *Real results, remaining challenges: The story of Texas educational reform*. The Business Roundtable.

Kelley, C., & Odden, A. (1995). *Reinventing teacher compensation systems.* Madison, WI: Consortium for Policy Research in Education.

Lewis, A. C. (2001). A performance test for districts and States. *Phi Delta Kappan, 82*(8), 567-578.

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems.* Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Marzano, R. J. (2000). *A New era of school reform: Going where the research takes us*. Aurora, CO: Mid-continent Research for Education and Learning.

Mathews, D. (1996). *Is there a public for public schools?* Dayton, OH: Kettering Foundation Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education, Macmillan.

Pan, D. T., & Mutchler, S. E. (2000). *Calling the roll: Study circles for better schools.* Austin, TX: SEDL.

Popham, W. J. (1999, March). *Why standardized tests don't measure educational quality [electronic version]*. Retrieved October, 2002, from www.ascd.org/readingroom/edlead/9903/extpopham.html

The Princeton Review. (2002). *Testing the testers 2002: An annual ranking of state accountability systems.* New York: The Princeton Review.

Ravitch, D. (2000). *Left back: A century of failed school reforms.* New York: Simon & Schuster.

Reeves, D. B. (2002, March/April). *Accountability-based reforms should lead to better teaching and learning — period*. Retrieved October, 2002, from www.edletter.org/current/reeves.shtml

Scheurich, J. J., Skrla, L., & Johnson, J. F. (2000). Thinking carefully about equity and accountability. *Phi Delta Kappan, 82*(4), 293–299.

Sirotnik, K. A., & Kimball, K. (1999). Standards for Standards-based Accountability Systems. *Phi Delta Kappan, 81*(3), 209–214.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oakes, CA: Sage Publications.

Walberg, H. J. (2002). Principles for accountability designs. In H. J. Walberg (Ed.), *School accountability* (pp. ix, 198). Stanford, CA: Hoover Institution Press, Stanford University Press.

APPENDIX A. SYNTHESIS OF RECOMMENDATIONS/GUIDELINES FOR ACCOUNTABILITY

Synthesis	CRESST	EdWeek	Princeton Review	RAND	Reeves	Sirotnik & Kimball	Walberg	WestEd / MAP
Clear standards and expectations	Accountability expectations made public and understandable to all. Stakes for results and phase-in schedule made explicit at outset.	Clear and specific standards	Are standards granular enough that a small number of items can reasonably measure a student's mastery of that granule?	Establishing clear teaching objectives by grade through statewide learning standards		System … must be based on high-quality content standards.	Focus on results	Student performance standards and aligned assessmer
High-quality assessments aligned with standards	Should document relationship between test items and specific standards Test results should be modifiable by quality instruction and student effort. Should document validity of tests for students with different language backgrounds and disabilities.	Criterion-referenced assessments aligned to state standards & external alignment reviews	Is there substantial overlap between standards and those that are actually tested? Are test items well written and the tests scored accurately and completely? Items validated before test construction?	Implementing new, statewide assessments closely linked to the learning standards	Bell curve is an ineffective, inappropriate way to measure achievement.			Alignment of state and local content standards
Multiple measures	Accountability systems should employ different types of data from multiple sources. Decisions about individual students should not be made on the basis of a single test.	Multiple forms of assessment/ accountability. State uses other forms of information to evaluate schools.	Do state report indicators of school quality other than static test scores, such as dropouts, teacher quality, or crime rates? Does it [report] judge schools by multiple measures?			The accountability system must not be driven by a single indicator (e.g., test scores) and simplistic formulas for rewards or sanctions based on that indicator.		
Public friendly	System results should be made broadly available to the press with sufficient time for reasonable analysis and with clear explanations of legitimate and potential illegitimate interpretations of results.	State requires that school report cards be sent home. State provides public with data on similar schools School report cards include ratings.	Are performance data shared with the public along with explanation and contextual detail appropriate for a general audience? Does state maintain publicly available data for evaluation of educational progress?				User friendliness. "Readily understood reporting is desirable."	

Synthesis	CRESST	EdWeek	Princeton Review	RAND	Reeves	Sirotnik & Kimball	Walberg	WestEd / MAP
Diagnostic uses	Reports to districts and schools should promote appropriate interpretations and use of results. Should include data that allow for interpretations of student, institution, and administrative performance.		Are complete test scores released to the public in a timely manner? Are test data distributed to educators in useful detail than can be linked to other databases?	Establishing a computerized system of feedback on test score performance at the student, classroom, school, and district level that can be used for diagnostic purposes	Once-a-year feedback is not sufficient. We should build system that gives monthly feedback to our children, our leaders, and our teachers.	Must include support for and monitoring of substantial, long-term professional development opportunities for teachers to review and revise their pedagogical content knowledge and teaching and leadership skills.	Timeliness. "One mark of a good teacher is getting the test back the next day."	Ongoing data analyses and reviews of school performance
Sanctions & rewards	Stakes in accountability systems (or incentives and sanctions) should apply to adults and students and be coordinated to support system goals. Should begin with broad, diffuse stakes and move to specific consequences for individuals and institutions.	State holds schools accountable for performance. System includes sanctions and rewards.		Establishing a system of accountability with both sanctions and rewards linked to the assessment results			Simply publishing results appears insufficient for progress.	
High expectations for all students (focus on at-risk students)	Accountability systems should include the performance of all students, including subgroups that historically have been difficult to assess.	Report cards include disaggregated data.	Are all students tested and included in statistical profile of the school? Does state publish detailed information on the performance of different groups?	Emphasizing strongly that all students are expected to meet the standards		Must include monitoring of and support for equitable and substantial learning opportunities for all students.		
Flexibility			Policy: Accountability system is consistent with state goals (flexibility)	Deregulating teaching & school environment; giving teachers and administrators more local control and increased flexibility in determining how to meet standards	Embrace different strategies so long as you report them.	The accountability system must be flexible enough to allow for individual differences in pace and style of learning — not a "one-size-fits-all" philosophy.		
Resources, support, & assistance		Assistance — State finances remediation for struggling students (ungraded)		Explicit shifting of resources to schools with more disadvantaged students		System must not be punitive;… must nurture and support struggling districts and schools. Must compensate educators at levels commensurate to critical importance of their work. Pay for it, or don't do it.		School improvement and intervention strategy

Synthesis	CRESST	EdWeek	Princeton Review	RAND	Reeves	Sirotnik & Kimball	Walberg	WestEd / MAP
Comprehen-siveness	Stakes for accountability systems should apply to adults and students and should be coordinated to support system goals. Asymmetry in stakes may have undesirable consequences, both perceived and real.				Let's focus on behaviors, not just test scores — in other words, measure what grown-ups do. We need to set as many standards for the adults — the board members, the administrators, the teachers, and perhaps someday even the parents — as we do for the kids.			
Stakeholder engagement / support						Public and political infrastructure must support the accountability system. Goals of system and the funding required to pay for it must have the support of the public and of the political infrastructure.		Stakeholder involvement and engagement
Fairness	Appeal procedures should be available to contest rewards and sanctions.		Do students have the opportunity to retake the test, if necessary? Are there due process guidelines for people accused of cheating?					