May 2000

Putting Standards to the Test: A Design for Evaluating the Systemic Reform of Education

Mike Puma
Urban Institute Education Policy Center, mpuma@chesapeake-research.com

Jacqueline Raphael
Urban Institute Education Policy Center

Kristen M. Olson
University of Nebraska-Lincoln, kolson5@unl.edu

Jane Hannaway
Urban Institute Education Policy Center

Putting Standards to the Test: A Design for Evaluating the Systemic Reform of Education

May 30, 2000

Mike Puma
Jacqueline Raphael
Kristen Olson
Jane Hannaway

Submitted to:

Policy Studies Associates
1718 Connecticut Avenue, NW
Washington, DC 20009

Prepared for:

US Department of Education
Planning and Evaluation Service
Collette Roney, Project Monitor
400 Maryland Avenue, SW
Washington, DC 20202

Prepared by:

Urban Institute
Education Policy Center
2100 M Street, NW
Suite 500
Washington, DC 20037

This work was supported by Contract EA9405301 under subcontract from Policy Studies Associates.

Chapter I: Systemic Reform — The Theory

Introduction: A Silent Revolution in American Education

A silent revolution has transformed American education rivaling the Progressive movement of the late 19th and early 20th century. This new movement — which goes by the name of "systemic" or "standards-based" reform — now dominates education policy in nearly every state, and is the basis for essentially all federal policy-making targeted at K-12 schools. This is not to ignore other competing strategies for the reform of our schools, most notably the inclusion of market competition (e.g., charter schools, school choice, vouchers), but even when these options are implemented they are typically being implemented within the broader context of systemic school reform.

The spark that ignited this revolution came not from national leadership but from the states, especially the highly influential report by the National Commission on Excellence in Education, *The Nation at Risk* (1983), which galvanized the education community around the goal of combating the "rising tide of mediocrity" that was purported to be destroying our schools and placing our children at risk of falling behind in the global marketplace. But even before the publication of this landmark report, a number of governors, especially those in the South, had placed education high on their political agenda based on a realization that without efforts to upgrade the skills of their state's children they would not be able to sustain economic growth in the new information-driven post-industrial economy.

By the end of the 1980's the governors were rapidly moving ahead with education reform and were then joined by President Bush in a 1989 "education summit" that created the first national education goals. A year later, President Bush proposed national legislation — the America 2000 Act — to implement the education goals using four strategies: (1) building local organizations to help achieve the national goals, (2) the design and implementation of "break the mold" schools, (3) demonstration grants to support school choice through tuition vouchers, and (4) voluntary national tests for grades 4, 8, and 12. It

was the issue of national testing, however, that eventually defeated this proposal in Congress.

Following the 1992 election, President Clinton (a former governor who helped lead the creation of the national education goals) made another attempt at stimulating the federal role in school reform with the introduction of "Goals 2000: Educate America Act," and proposals for incorporating systemic reform into the reauthorization of the Elementary and Secondary Education Act of 1965 (called the "Improving America's Schools Act" of 1994). In both cases, these legislative proposals represented dramatic departures from previous federal education policy. For more than three decades, the main purpose for federal intervention in locally-controlled public schools has been to promote an "equality of educational opportunity" for various groups of disadvantaged children. In 1965 Congress passed the initial landmark Elementary and Secondary Education Act (ESEA) which extended federal support to help disadvantaged students; this was followed by efforts targeted at bilingual students in 1968, the disabled in 1974, and minority students in 1983. Since the early 1990's, however, the thrust of federal policy has shifted more to the use of federal funds to encourage states to make broader changes in school systems. Most recently, we have seen this effort extended even farther with the use of federal funds to hire teachers, reduce class size, and increase school access to educational technology — all actions which are not necessarily targeted at specific groups of children.

The main change in the shift from previous federal legislation and the "America 2000" and "Goals 2000" programs was to a greater focus on state leadership as the driver of school reform, and the need for aligned and coherent policies regarding standards for what students are expected to learn, instructional materials and curriculum, teacher preparation, and accountability and assessment systems. The basis for this changed direction was derived in large part from the seminal work of Marshall S. Smith (later to become the Assistant Secretary of Education) and Jennifer O'Day, then at Stanford University. The legislation, enacted in 1994 after a year of continuing debates, acknowledged the role that many states were already playing in systemic reform. The

legislation support states in focusing more on the outcomes of district and school efforts (e.g., student achievement and changes in instruction) and less on compliance with rules and regulations. Specifically, states were encouraged to develop content and performance standards in core subject areas, and to align their entire educational systems — including assessment, curriculum, instruction, professional development, and parental and community involvement — around these standards. As a result, Goals 2000 began serving as a source of funds that could support state efforts already underway, as long as they conformed to the general principles of systemic reform.

Since passage of Goals 2000, Congress has appropriated over $2.5 billion for this purpose, with at least 90 percent of each state's award subsequently sub-granted to school districts and/or consortia of districts (after year 1, when 60 percent was sub-granted) to implement systemic standards-based reform efforts at the local level. According to a 1998 Government Accounting Office report, between 1994 and 1997, states made Goals 2000 sub-grants to over one-third of the 14,367 districts in the Nation. Generally, funds have been used to support state reform efforts through: the development of state and local standards, curricula, and assessments; professional development related to the new standards and curricula; and, improved pre-service teacher education.

This chapter presents the theory underlying the systemic reform movement in education. The next section details Smith and O'Day's approach to systemic reform, forming the theory behind the Goals 2000 legislation. The following four sections examine systems theory, how K-12 education in the United States functions as a system, and two conceptual models of how the U.S. K-12 education system operates under systemic reform. The chapter concludes with a brief critique of systemic reform and a rebuttal to this critique.

The Seminal Work of Smith & O'Day

As noted above, the seminal work of Smith & O'Day (1990) helped spur this new wave of school reform by offering a new approach to thinking about how to improve American education. In their view, past "waves" of school reform (e.g., tougher state graduation

and promotion requirements, site-based school management) had failed because schools were prevented from being successful by the fragmented and multi-layered educational policy systems in which they were embedded. What was needed, according to Smith & O'Day, was a fundamental restructuring of instructional content, pedagogy, and the overall educational system:

> *…if we are to significantly alter student outcomes, we must change what happens at the most basic school level of education — in the classrooms and schools. However, we see in this process a more proactive role for centralized elements of the system — particularly states — one which can set the conditions for change to take place not just in a small handful of schools or for a few children, but in the great majority.* (p. 235)

What made this idea so radical was that it sought to fundamentally alter the way educational changes were traditionally made by moving from an *incremental* approach — adjusting a single component of the instructional process such as reducing class size — to a *systemic* perspective, in which reforms should seek to change how the different components of an educational system worked together. The underlying hypothesis was that increasing the alignment or coherence among the different components, actors, and agencies that make up the complex enterprise of education (i.e., increasing the "system-ness"), would make schools more effective, and this, in turn, would improve the system's "output," i.e., teaching and learning.

In explicating their new vision, Smith & O'Day focused on several important ingredients in key areas for their recipe of systemic school reform:

1. Governance

- **A Unifying Vision and a Consensus on Values.** To create a coherent system requires a common vision about a good school including "…a schoolwide vision and school climate conducive to learning, enthusiastic and knowledgeable teachers, a high quality curriculum and instructional strategies, a high level of engagement, shared decision making, and parental support and involvement" (p. 236). The type of broad change envisioned by the authors also required a consensus on key values, including "…respect for all people, tolerance, equality of opportunity, respect for the individual, participation in the democratic functions of society, and service to the society……(and) to prize exploration and production of knowledge, rigor in thinking,

and sustained intellectual effort" (p. 246). These visions and beliefs must, in the authors' view, permeate the entire educational system.

- **Overall State Policy Leadership.** The authors focused their primary attention on the role of the state in the process of driving broad systemic reform. In their view, "top-down" leadership is needed because: (1) only states can create change in many districts and schools rather than for just a few at a time; (2) states have gained increasing authority over education due to growing concerns about the importance of education for economic productivity, and the need to ensure equity in the access to resources and services; and, (3) only states are in the position to influence all parts of the educational system from curriculum, to teacher training and licensure, to assessment and accountability. They also recognized, however, that states are not monolithic, and ultimately, success in systemic reform would require cooperation among the various key state-level players including the state superintendent, the governor, and the legislature.

- **The "Bottom Up" Role of Districts and Schools.** Despite the prominence given to the role of states, Smith & O'Day recognized the need for "bottom up" involvement from district and school staff.

  Districts were expected to create policies that worked in concert with the direction set by their states, and yet were responsive to local needs and conditions. In particular, districts were expected to "provide resources and a supportive environment" that allowed schools to achieve their mission of educating students, including: reducing centralized bureaucracy; reducing policies and rules that inhibit innovation and effective school-based instructional approaches (e.g., rigid class size and time requirements, conformity in textbooks and instructional materials, etc.); and ensuring the equitable treatment of all students. Achieving these objectives, according to Smith & O'Day, requires that district administrators, school boards, and unions work "toward strategies that ensure policy continuity rather than disruption," and that avoid ineffective attempts at "quick fixes." The authors also recognized that "…many districts will have difficulty in altering their procedures and modes of behavior……in some cases the talent is not presently available…..in other instances the central administration is simply resistant to significant change" (p. 257).

  Schools were expected to create "…a stimulating, supportive, and creative environment to maximize student achievement" (p. 254). This includes: a positive atmosphere and a high level of respect among students and staff; well-trained professionals; school-based goals; empowered teachers who are in decision-making roles regarding the design of instruction; adequate time for planning, collaboration, and professional development; flexible organization of student time (e.g., small groupings, flexible time allocations, cross-age tutoring, cooperative learning, etc.); mechanisms for parental involvement; and effective use of educational technology and resources to support instruction and how teachers work.

2. Alignment of Curriculum and Instruction

The articulated vision and values should support a coherent curriculum and instructional framework. According to Smith & O'Day, these require: (1) agreement on what students should know and be able to do , and (2) assurances that students have the opportunity to acquire these competencies by being exposed to the requisite knowledge and skills. Further, to accomplish these objectives requires coordination among state standards, school-level curricula, pre-service and in-service professional development, teacher certification, and statewide assessment and monitoring:

- **Measurable Goals** — Smith & O'Day emphasized the need for educational goals that can be communicated and measured using a system of indicators "…that challenges the public and the educational system to prepare our youth ….to be skilled and confident learners in school and later on. Moreover, the goals and indicators must address not only the average level of opportunity and student achievement in the state but also the variation. Justice requires that the goals of the state promote equality as well as quality" (p. 247). Such goals and indicators, according to the authors, are needed to mobilize and sustain political support, and to provide those within the system, and the public, with a sense of direction and a basis for monitoring progress. Some indicators should address changes in the quality and nature of educational inputs (e.g., teacher skills), and others should measure student outcomes: school readiness, self-worth, academic achievement.

- **State-level Curriculum Frameworks** — Curriculum frameworks —now typically called "content standards"[1] — are described as "roadmaps" that allow for local flexibility and innovation with regard to what should specifically be taught and how. As such, they specify the knowledge, processes, and skills that K-12 students are expected to know in core subjects of "…reading/language arts, English, math, science, social studies and history, foreign languages, and the arts." These frameworks are, however, expected to go beyond "fact-based" education to "…emphasize depth of understanding, knowledge construction through analysis and synthesis of real-life problems, hands-on experiences, and the integration of content and pedagogy" (Smith & O'Day, 1990).

  In effect, the frameworks are expected to serve as "…a structure within which to organize the other important educational components." This means that they should be aligned with: professional development (pre-service and in-service) and licensure requirements to ensure that teachers are well prepared to teach the specified content; curriculum, textbooks, and instructional materials; and, tests and assessments "…used to assess pupil progress and to hold schools and teachers accountable." The authors

[1] Smith & O'Day (1990) used the term "curriculum framework" to describe what currently is typically called a state's " standards." Curriculum frameworks, as the term is used today, establish a bridge between state content and performance standards and local curriculum, providing guidance on curriculum, teaching strategies, use of technology and other materials, and assessments. Forty-one states have or were developing state curriculum frameworks, according to the Council of Chief State School Officers' (1999) and a review of state department of education web sites. Often these frameworks contain (and therefore are indistinguishable from) the state content standards.

also recommended that the frameworks should be: (1) developed by "highly qualified teams of teachers and disciplinary experts"; (2) "continually updated and reviewed by similarly qualified expert panels"; (3) of the highest quality to command the respect and enthusiasm of capable teachers; and, (4) flexible to allow local school personnel the freedom to interpret and implement instructional strategies that most effectively meet the needs of their students.

- **School Curricula** — As noted above, Smith & O'Day did not envision a solely "top-down" model of school reform. Rather, they emphasized the need for schools and districts to have the flexibility and support to construct strong locally-responsive curricula within the state curriculum frameworks that are "…best suited to their students and teachers."

- **Professional Development** — according to Smith & O'Day, "States must ensure that both new and practicing teachers have the content knowledge and instructional skills required to teach the content of the frameworks." In other words, the frameworks are intended to define not only what students should know, but what teachers should know and be able to teach. This means affecting both pre-service and in-service training:

  - *Pre-service professional development* — Over time, the authors believed that the development and establishment of curriculum frameworks would help push schools of education to meet higher standards and requirements for prospective teachers.

  - *In-service professional development* — Even with this change, the authors emphasized the need for continuing high-quality in-service professional development both as a way to compensate for the inadequate training of many teachers, and as a way to "empower the teaching force." States were expected to influence the supply of high-quality professional development programs and materials, to allocate resources for the development and implementation of such programs, and/or to provide incentives for schools and teachers to take part in training programs. In particular, states were expected to motivate teachers and other staff to want to improve their knowledge and skills by holding teachers and schools "accountable for improving student outcomes on assessment instruments that are based on the frameworks," or through the use of state licensure exams tied to the frameworks.

3. Accountability and Incentives

In addition to supporting efforts to increase the "capacity" of the educational system (through professional development), states were expected to also send strong "signals" regarding the need to improve educational outcomes.

- **Aligned Assessments.** The authors further recommended the construction and administration of "…high quality assessment instruments on a regular basis to

monitor progress toward achievement goals for accountability purposes and to stimulate and support superior instruction" (p. 252). Such assessments would be tied to the curriculum frameworks, requiring a complete overhaul of existing testing systems so that the assessments truly provided a measure of what schools and teachers are expected to teach. Previously, many states used standardized, norm-referenced tests to measure student achievement. However, these tests were designed by test developers through a review of major textbooks throughout the country and did not adequately reflect the curriculum that was being implemented in specific schools. The aligned assessments would be content-driven and based on the expectations set out in the state curriculum frameworks. These new assessments would have to "…encourage instruction toward higher level goals: depth of knowledge, complex thinking, an ability to respond to problems and to produce results" (p. 253). In this way, teachers and students would have a clear idea about what the system was "…striving for and a way to monitor success in getting there." To avoid overwhelming the schools with testing, Smith & O'Day recommended selecting a small number of grade levels (e.g., 4th, 8th, and 11th) to be the focus of the state assessments.

- **Accountability and Incentives.** Districts and schools should also be held "…responsible for demonstrating either an across-the-board high level of achievement for their students or a steady growth over time in that achievement. Assessment for accountability purposes could also be combined with incentive measures for meeting or surpassing objectives" (p. 253). One point the authors were uncertain about was the scope of such assessments. On the one hand, if district and school accountability were the goal, then only samples of students should be tested (thereby reducing some of the burden). Alternatively, if the assessments were also intended as a way to motivate students to study, then the tests would have to both be administered to all students and would have to involve "high-stakes," i.e., the results would have to involve consequences for students, not just for their schools.

- **Equity.** Finally, Smith & O'Day clearly recognized that one danger in such systemic educational reform is that poor and minority students could be left behind because poor districts and schools "…have less discretionary funds to stimulate reform, less well-trained teachers, and more day-to-day problems that drain administrative energy" (p. 258). The authors recommend that states ensure equity of opportunity; a failure to do so could expand the differences between advantaged and disadvantaged students and schools as part of a drive to create more challenging curriculum and instruction.

According to Clune (1993), systemic educational policy rests on four assumptions implicit in the work of Smith & O'Day about what ails the current K-12 public education system: (1) curriculum is generally poor and unchallenging — a more challenging curriculum would have a strong positive effect on student achievement; (2) a combination of coherent signals and capacity building from outside schools is needed to encourage

and sustain positive educational reform, especially at the school level; (3) the prevalent multi-level system of policies pushes the curriculum toward mediocrity characterized by fragmentation and contradiction; and, (4) there is a lack of well-accepted (and documented) challenging goals for student achievement.

The Systems Approach: What Is It? How Is It Different?

Probably the most important insight offered by Smith & O'Day, and which is at the root of the systemic, standards-based reform movement, is that attempts to change student educational outcomes must involve a systems perspective.[2] Webster defines a "system" as "a regularly interacting or interdependent group of items forming a unified whole" that are "under the influence of related forces" and which form "a network especially for distributing something or serving a common purpose (e.g., a telephone *system,* heating *system,* highway *system,* data processing *system*)." Embedded within the system definition is the concept of "order, i.e., a harmonious arrangement or pattern — to bring *system* out of confusion."

Traditionally, the approach of finding solutions to education and many other public policy questions has been based on the use of an *analytical* lens to break problems into small component parts that can be separately studied (i.e., defining their static characteristics). This approach, which has dominated scientific and social discourse since the 16th and 17th centuries, is based on the theory that by knowing the nature of the parts (e.g., individual businesses or people) one can understand the larger collective organization (e.g., an economy, organization, or society).[3]

Instead, the systems approach emphasizes the study of interactions and connections among the interdependent components, i.e., how order is derived from a network of separate parts. Systems theory, which emerged after World War II with the initial work of

[2] For this discussion we are focusing on public K-12 education.

[3] Similarly, this approach has dominated efforts to reform education prior to standards-based reform, with curriculum reform (e.g., the "New Math" curriculum of the late 1950s) viewed as the "silver bullet" for large-scale, far-reaching change.

Von Bertalanffy and others,[4] defines a system as an organized whole which is composed of two or more inter-dependent parts (sub-systems) and which has identifiable boundaries separating it from its environment (the supra-system). A tree is an example of a system that is comprised of separate systems (e.g., a root system, a trunk serving as a structural support system, etc.), and which is clearly separated from its environment (i.e., the ground, air, other organisms) with which it interacts. A key characteristic of this view is that a system must be viewed as a whole — changes in one part of the system affect all the other parts.

Types of Systems

Systems can be either "closed" (they do not interact with their environment, e.g., a clock), or "open," in which case there is an important interaction or dependency between the system and its environment (e.g., humans). The important characteristic of open systems is the complexity introduced by changes over time as the system evolves new functional structures in response to interactions with the environment, or even "self-organizing complexity," in which the system co-evolves with the environment (examples include natural ecological systems and languages).

Mutually evolving open systems, also referred to as "complex adaptive systems," share a variety of characteristics. First, they typically consist of a network of components — called sub-systems — that constantly affect each other. Second, these sub-systems are not centrally controlled. Instead, patterns of form and function arise as a consequence of the interactions that occur, with each component continually affecting, and being affected by, the other sub-systems. Third, the sub-systems are often organized into different "levels," in which lower levels serve as building blocks for higher levels (e.g., workers forms teams that then form larger organizations). Finally, these systems constantly re-organize themselves as functions and structures change over time.

[4] See, for example, some of the classic works in the field: Von Bertalanffy, L. (1976). ***General Systems Theory***, George Braziller; Weiner, N. (1986). ***Norbert Weiner: Cybernetics, Science, and Society (Collected Works)***. MIT Press; and Forrester, J. (1968). ***Principles of Systems.*** Pegasus.

Parts of Systems

As noted above, systems are made up of "sub-systems" that are in turn comprised of "components" which can be part of more than one sub-system. Examples of different types of sub-systems comprising a complex adaptive system include the following:

- ***Production and technical sub-system*** which is responsible for converting inputs (e.g., teachers, instructional materials, etc.) to products or services (e.g., teaching and learning).

- ***Supportive sub-system*** is responsible for two major function, (a) obtaining inputs, and (b) promoting and maintaining good relationships with the environment (e.g., parent and community relations).

- ***Maintenance sub-system*** deals with personnel (e.g., teachers) in the system including their recruitment, selection, role assignments, and motivation needed to maintain stability in the system.

- ***Adaptive sub-system*** involves functions necessary to ensure that the system can meet the changing needs of the environment. This includes tasks such as planning, and research and evaluation.

- ***Managerial subsystem***, this final sub-system coordinates the function of all other sub-systems, settles conflicts among them, and relates the system to its environment. Its cross-cutting functions create the alignment and coherence of action necessary to achieve the highest levels of performance.

Important building blocks of all systems, but especially complex adaptive ones, are "links" and "loops" — connections that tie sub-systems together. These can be of two types: *reinforcing*, where small changes or disturbances can lead to exponential growth or decline (e.g., a small change in a birth rate can lead to rapid population growth); or, *balancing*, where there are inherent forces of correction or resistance that maintain stability or equilibrium (e.g., a self-correcting or regulating control such as a thermostat that maintains a constant temperature). *Delays* or lags can, however, occur in both types of loops where the chain of influence takes a long time to play out. In balancing loops, for example, delays can create wild oscillations as reactions are out of proportion with the initial need for change.

Information Flow In a System

An important way to study systems is to map the pathways of information flow (or alternatively, the flows of energy, matter, etc.[5]). Called "cybernetics" this view of complex systems focuses on communication and control interactions (i.e., exchanges of information) both between the system and its environment, and among sub-systems, using loops and links[6]. In other words, this framework describes how systems get information, how they incorporate information to understand their surroundings, and how they make decisions based on the information. This is what Simon (1969) calls "blueprints" — descriptions of state — and "recipes" — prescriptions for action. Systems learn about their environment by attempting to control it, and modify their representation of the environment by monitoring the results. This process is often referred to as feedback.

Traditional systems theory focuses on the **structure** of systems and attempts to create models of how they work. In contrast, cybernetics focuses more on how systems **function,** that is, how they control their actions and communicate with other systems, or with their own sub-systems. Of course, structure and function are two sides of the same coin and are often inseparable. This idea of communication, control, and feedback is critical to an understanding of systems. The example shown in Exhibit I.1 (see next page) can help explain how this works.

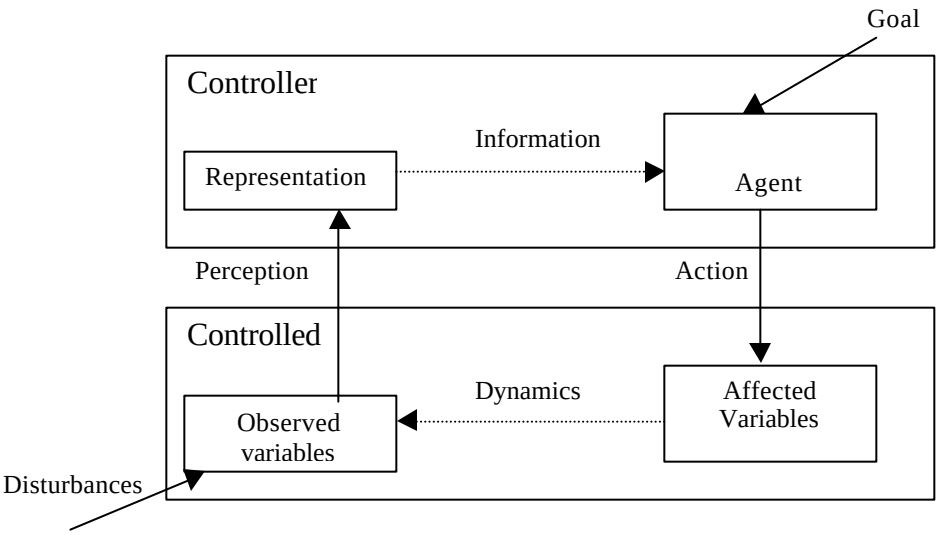
In this depiction, the **agent** (the controller) seeks to achieve some **goal**. To do so involves taking an **action** on the **controlled** sub-system. The controlled system is described using some variables and distinguishing between the variables that are directly **affected** by the controller, from the variables that are **observed** by the controller in **perception**. The causal relationship between the observed and affected variables is determined by the **dynamics** of the system, and uncontrollable **disturbances**. The agent compares the

[5] A unique attribute of information is that while it can move from A to B, a copy can remain at A.

[6] The term cybernetics — a fundamental part of complex systems theory — is, in fact, derived from the Greek word (kybernetes) for steersman or pilot, i.e., one who controls.

current **representation** with the goal and takes further actions to minimize the difference between them. From the controller's perspective, this "loop" begins with an action, followed by a perception of the results of that action, and continued action to get closer to the achievement of the goal (called **purposeful behavior**). The perception is an action in the opposite direction (from the controlled to the controller) and called **feedback**. It is this set of relationships that form the building blocks of complex systems.

Exhibit I.1: Example of Control and Communication in Systems



K-12 Education as a System

With this theory as background, how can we begin to think about the education as a system as described by Smith & O'Day? This is important to do because some have criticized their work, claiming that no scientific basis exists for a theory of systemic reform. According to Bruckerhoff (1997), "…there is only a strong belief, or rationale, that changing the system of education in accordance with these normative principles will lead to significant increases in student ... achievement." Sounding the same theme, Hatry & Kopczynski (1996), experts in performance measurement, agree that "a major obstacle ... for any effort to assess progress in systemic reform is the lack of clear definitions of what systemic reform and its components are." Weiss (1999), who has considerable

experience evaluating systemic reform initiatives, writes that "…systemic reform theory is exceedingly thin, specifying overall goals, but providing little guidance on how to go about meeting those goals." What these and other reviewers suggest is that although systemic reform has contributed to the debate about how to fix our schools, it is still in many respects a "work in progress."

To begin with, it is reasonably clear that education fits the model of a complex system made up of many different parts (e.g., teachers, curriculum, etc.), which are continuously interacting both with each other (e.g., teacher training affecting instructional practice) and with the outside environment, i.e., with other "levels" of the larger education system (preschool/early childhood education, post-secondary education, "life long" learning), and with the broader social and political system. These interactions also create an adaptive system (i.e., responding to changes based on the results of actions) that is co-evolving with its environment (i.e., responding to changes in the types of students coming into the system, changes in available technologies, and changing demands for what it means to be "educated" in the current society).

Traditionally, analysts have tended to treat this complex system as a "black box" with inputs (e.g., students and resources) at one end, and outputs observed at the other end (e.g., academic achievement measured by test scores and rates of promotion and graduation). This approach is exemplified by the widespread use of econometric or statistical models of the "education production function." In contrast, a systems view is a "white box" perspective that tries to understand the internal processes that converts inputs to outputs. But, more importantly, it rejects the reductionist viewpoint that all you need to know is the precise state of the internal components to understand how the system functions. According to the system perspective, by ignoring "the whole" one fails to understand that complex systems are not a simple combination of the constituent parts, but instead a complex network of interdependencies. When we say that "the whole is more than the sum of its parts" it is this notion that the "more" refers to higher level laws which make the parts function in a way that does not necessarily follow from the patterns of behavior of any one component.

What this perspective emphasizes is the importance of the *relationships* among sub-systems rather than their operational characteristics, i.e., a focus on the links and loops, the patterns of communication and control, that create the "system" of education. This is what Smith & O'Day mean by the need to move away from the prevailing state of fragmentation and incoherence to a system of education that is characterized by the **alignment** of goals, policies, procedures, and capacity-building through the creation of a new **governance** structure, and maintaining the "fitness" of the system through the use of **accountability and incentives**.

A Conceptual Model — A First Cut

Exhibit I.2 is an attempt to depict K-12 education as a complex system of connected sub-systems that interacts with the outside environment and changes over time in response both to environmental demands and internal improvement. Within this simplified model,[7] we recognize five inter-related sub-systems:

- ***Policy Infrastructure***: This "managerial sub-system" operates primarily at the state level to manage and coordinate the operations of the other sub-systems. One side of this component deals with **policy instruments** including creating and maintaining common *goals and vision*, and developing *content and performance standards and expectations* for students. The other side of this sub-system is concerned with the **change strategy** that will guide reform including:

    1. **State/district inter-governmental relations** — how roles and responsibilities are divided across governance levels.

    2. **Sequencing** — decisions about the order in which components of reform should be implemented (e.g., should assessments start the process to provide strong "signals" to all the actors in the system, or should standards and frameworks be initially developed and the other parts designed to flow from them).

    3. **Communication and leadership** — how policy will be communicated throughout the system, and with "outside" parties, and the tools that are brought to bear to effect the desired changes.

[7] We have ignored many important sub-systems such as resource supply (books, instructional materials, technology), external communication (public relations), and planning (including R&D).

4. **Accountability** — the rewards and sanctions that will be used to help drive changes in performance.

5. **Practitioner involvement** — the extent to which, and how, district and school staff are involved in the creation of the state-level policy instruments.

- ***Curriculum***: This first "technical sub-system" is where the broad policies are converted into specific curriculum and instructional guidance that can actually be used to guide classroom instruction.

- ***Assessment***: This "adaptive sub-system" creates the information needed to determine how well the system is functioning through the creation of information and for continuous, data-driven improvement.

- ***Instructional Practice***: This second technical sub-system is the true work-horse of education — "where the rubber meets the road" — involving the delivery of curriculum and instruction at the classroom level.

- ***Capacity Building****:* This final "maintenance sub-system" is concerned with acquiring, retaining, and building the human capacity to deliver high-quality instruction. As a consequence, this sub-system includes staff recruitment and selection, pre-service training, professional development, incentives tied to performance as a way to sustain high-quality performance, and the use of data for continuous organizational improvement.

The model ends with a box containing system outcomes — student academic achievement and equitable opportunities for *all* students. In addition, the **environment** presents important contextual factors that affect the implementation and outcomes of systemic standards-based reform: parent/community involvement in education, including parents' input about education into their children's lives; changes in the student population (e.g., increasing diversity and high student mobility); the skill/ability needs for which schools prepare students; changes in non-K-12 education for which students may be eligible in the future (e.g., higher education, vocational education, adult education); the politics of the community; as well as other factors. Finally, we include the evolution of the system over time, to acknowledge that this is an adaptive system which is expected to change over time in response to internal and external demands.

The different sub-systems shown in this stylized model are all **autonomous** in the sense that each component "decides" its own actions. However, these sub-systems operate

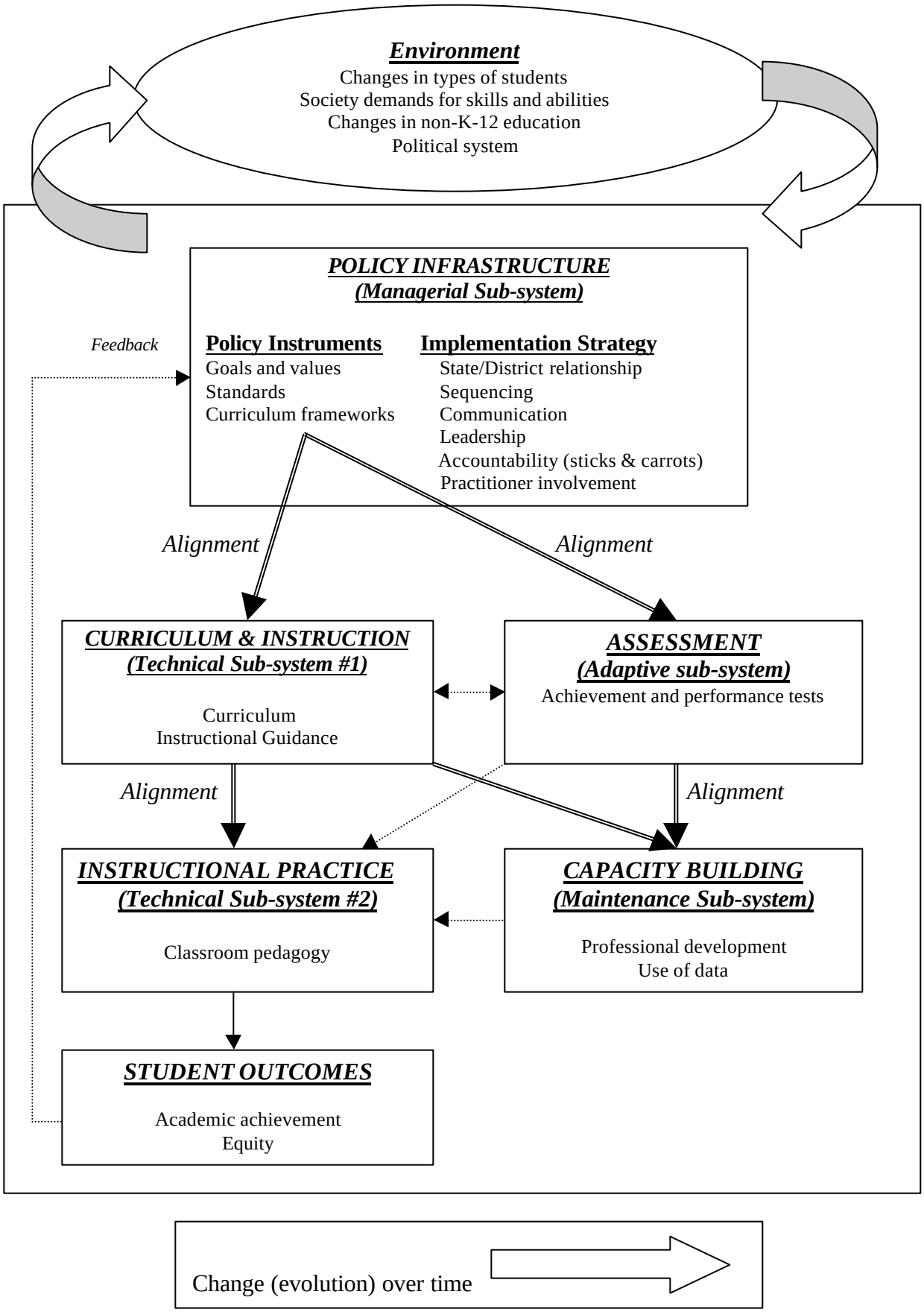


Exhibit I.2: Stylized Model of the K-12 Education System

within the constraints of the overall system. For example, teachers have a great deal of autonomy within their classrooms (the "instructional practice" box) but are constrained by a variety of school and district policies, practices, and organizational culture (i.e., informal rules). As a result, how the system functions is determined by the **pattern of interactions** among the various sub-systems, i.e., the dynamic relationships among the parts of the overall system are more important to its survivability and growth, and are also more germane to understanding "how it works" than knowledge of the workings of a particular sub-system.

In the theory of systemic reform the interactions that are of primary interest (highlighted in the exhibit) are those that create **alignment** or coherence between the overriding goals of the system (also expressed as content and performance standards) and the other parts of the K-12 educational system. That is, the greater the alignment, the greater the "system-ness" as defined by Smith & O'Day. In systems theory alignment is often defined in terms of the strength of the **coupling** that exists between and among the different sub-systems, i.e., the level of inter-dependence that exists. In particular, the thick lines with arrows in Exhibit I.2 indicate where alignment between the sub-systems is needed to ensure that the education system reforms are truly *systemic* and *standards-based*. If the coupling is too "loose," then information flows are weak, leading to the type of fragmentation and lack of coordination that Smith & O'Day (1990) point to in their article. At the same time, if the coupling is too "tight," then anything that happens in one sub-system will reverberate throughout the entire system (like a domino effect). The trick is to find the right degree of "coupling" that provides the necessary interaction without constraining the ability to respond to changes and innovations. For example, if standards and assessments are too prescriptive this could constrain innovation at the school and classroom level and make instruction less responsive to the needs of individual students.

To help illustrate the conceptual model, consider a hypothetical state and start with the top box labeled "policy infrastructure." Consider that this state wants to improve academic achievement Under systemic reform the state would attempt to achieve this objective through the use of **signals** represented by *standards*. These standards are, in

turn, converted to *curriculum and instructional guidance* that affect important aspects of instruction. How instruction is affected by these signals is subsequently observed, at the state and district level, through the use of student *assessments* that seek to measure the congruence between the standards and *student performance*, provided the assessments being used to measure student performance are well-aligned with the curriculum. (This is an important source of "feedback" in the system.)

But, as we know, what is observed through an such *accountability process* (e.g., student test scores) is not directly caused by the selected *policy instruments.* What we see is instead the result of a complex dynamic process that includes the translation of the frameworks into curricula, and then into actual classroom practice, and finally into student learning. In large part, how well this process works is related to the *environment* and the *implementation strategy* that is chosen, and is also affected by a variety of intervening disturbances (e.g., random variation in teacher skills). As a consequence, the state's perception of "what happened" as a result of the promulgation of the curriculum frameworks typically is not exactly in line with the original goals that were expected to be achieved; it reflects the influence of several intervening variables.

Those involved in instructional practice (e.g., teachers and school administrators) are not, however, passive agents in this process. They also send signals (information) back regarding how well the standards are actually working in practice (a second source of feedback). This combination of feedback flows allow the state to assess the extent to which the result is at variance with the original goal, subject to factors that may affect interpretation of the information (e.g., political pressures), and this then triggers some further action intended to close the perceived gap. Depending on the strength of the "coupling" that exists among the various sub-systems the process of attaining the desired goal will be a smooth or bumpy ride, characterized by wild swings or small adjustments in policy and practice.

The same sort of interactions — information flows and feedbacks — exist among the other sub-systems but of course their nature will differ depending on the sub-systems

involved. Of particular importance within the education context is the distinction between "signals" and "capacity development." The establishment of goals and standards, and accountability systems, are both system signaling procedures that are expected to cause intended actions on the part of the other sub-systems. For example, one of the assumptions in the work of Smith & O'Day is that by setting expectations and measuring performance against those goals, teachers, administrators, parents, and students will be "spurred" to higher levels of instructional quality and learning. The authors, however, acknowledge that this is likely to be insufficient, especially in the most disadvantaged schools, where staff may lack the support and resources necessary to meet the expected standards. As a result, other sub-systems (primarily the "capacity building" sub-system responsible for training and motivating staff) are expected to provide a different type of information to the technical sub-system in the form of guidance, mentoring, or coaching, among other forms, that will help build the capacity required to adequately respond to the policy signals.

A Second Perspective

Exhibit I.2 is but one view of the education system because it ignores the various levels of political governance that currently exist in the US. A way to think about this expanded perspective is to see the "first cut" model as a cube rather than as a flat plane. In this refined 3-dimensional perspective, Exhibit I.2 is but a single view through one surface of the cube, i.e., along dimensions characterized by system processes. If we now rotate the cube and peer through an adjoining surface we will see that each process sub-system consists of levels representing the state, district, school, and classroom "building blocks" that comprise the K-12 education system.

This additional perspective, shown in Exhibit I.3, depicts the **holarchic** relationships among the different governance levels as they relate to curriculum and instruction (i.e., the levels are depicted as being contained within each other instead of hierarchically dominating each other). In this simplified model, the role of the state is shown as developing the curriculum frameworks in response to overall goals and agreed upon values for K-12 education. The state also continues to operate within its environment

responding to external constraints, pressures, and changes that affect the way in which it operates and the types of decisions that are made regarding standards. Yet states are not expected to operate in isolation, but are instead supposed to develop standards using input from district and school practitioners, and this is expected to be a dynamic process helping to evolve the standards and frameworks over time. These relationships are represented by the feedback loops among the different levels.

School districts are, in turn, expected to "translate" the curriculum frameworks into locally relevant and responsive curricula and instructional guidance, also with the input of actors in the next level of the system (schools), and within the context of their local environment, i.e., they are also "open systems" that are affected by external stimuli including changing student body composition, administrative and political pressures, and parent and community influences. Further, schools and teachers (the classroom in this model) are the "front line" of the overall system and are expected to implement the curriculum through the application of pedagogical skills, and an assembly of school and classroom routines and management techniques. Moreover, schools are also open systems and must do so within the context of their particular local environment.

These collective school, classroom, and community experiences and knowledge are, according to Smith & O'Day, expected to flow back through the different levels in what they call a "top-down, bottom-up" chain of interactions that help to improve and enrich the entire process of systemic change.

The Voice of the Critics

Since the publication of the Smith & O'Day paper there have been countless articles and commentaries published, many of them critical of this theoretical perspective for a variety of reasons. For example, the National Academy of Education's Panel on Standards-based Reform (1995) raised several concerns about the setting of common standards: the focus on results could deflect attention from delivery issues and other problems in education; uniform content standards could lead to a narrow curriculum that

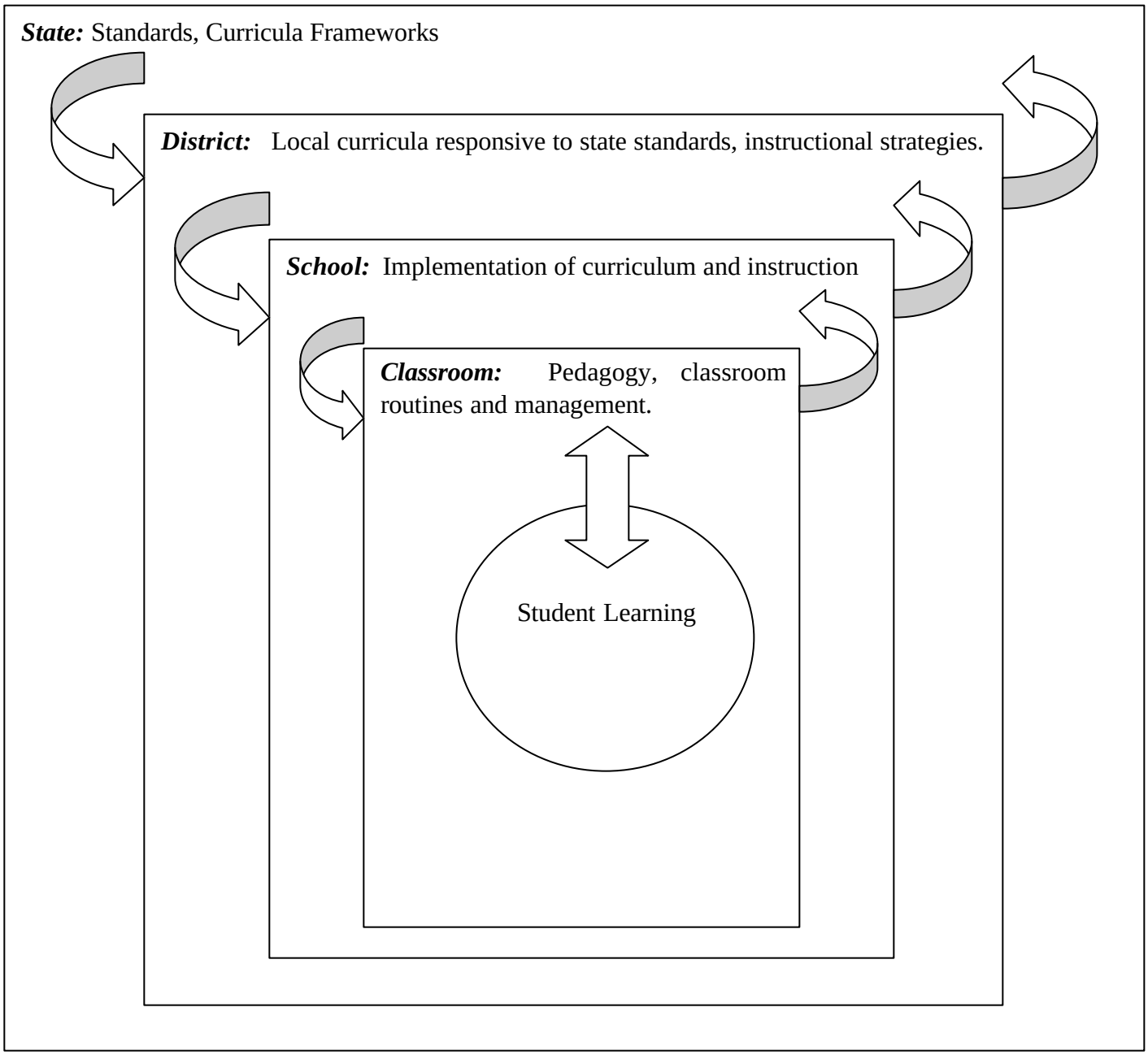


Exhibit I.3: Incorporating Governance Levels: Curriculum and Instruction

is inconsistent with constructivist theories of knowledge; and, standards linked to assessments may contribute to greater educational inequities if teachers in low-income schools are not provided with the support needed to shift to more reform-oriented pedagogy.

Others have focused on the governance features of the theory, in particular the balance of power between states, districts, and schools. For example, Wang, et al. (1993) argue that many of the recommended systemic changes have not been found to exert a significant influence on student learning, in large part because broad policy changes do not reach down to the classroom level to have a significant effect on what teachers do in their classrooms and what students subsequently learn. Others have suggested that increased centralization could potentially limit local innovation, teacher responsibility, and sensitivity to local educational needs (Knapp, 1997). Further, Clune (1998) argues that a centralized systemic education policy is infeasible because "…a common curriculum is difficult, if not impossible, to apply considering the immense diversity of American schooling."

A number of other critics have reacted to Smith & O'Day's lack of emphasis on the district as a significant "player" in reform efforts. Mitchell & Raphael (1999), for example, demonstrate the essential role of districts in school reform especially their ability to introduce school-level officials to new instructional strategies. This finding is corroborated by Fuhrman, Clune, & Elmore (1998) who found that districts are able to mobilize local support for schools, may serve as a resource for schools in terms of innovative school-reform ideas, can bring a broad base of stakeholders to the table, and can serve as needed administrators between state policy and school-level reality. Similarly, Fullan (1994) and other researchers cited by Fullan note that change is most likely to occur in systems where district- and school-level initiatives are undertaken simultaneously in a process of "co-management." Seen in this light, one could argue that districts do not simply mediate state policy, but can transform it into coordinated local strategies that produce school change (Spillane, 1998). Indeed, exemplary districts appear

to respond in a proactive manner to both state- and school-level reform activities and policies (Elmore, 1998).

Finally, critics such as McLaughlin & Shepard (1995) have raised concerns about the potential for a standard state curriculum to undermine professional and local responsibility for student learning, even when such an approach is discouraged by federal agencies. That is, can the public education system use standards as tools rather than as mandates for instruction? If students learn by constructing knowledge then teaching must be adaptive, able to elicit relevant prior knowledge, attuned to local contexts and experiences, to make connections, and open to emergent understandings expressed by students.

Is This Criticism Warranted?

In most cases, these varying criticisms are speculations about what "might" happen (e.g., empirical questions about the potential for negative consequences for poor children, or a loss of local flexibility), a misreading of the original theory (e.g., incorrect claims that Smith & O'Day ignore the role of the district in systemic reform), or an attack on **parts** of the theory while ignoring the critical **systemic** nature of the original concept (e.g., focusing on the nature of standards, or the potential for abuse associated with the use of assessment and accountability systems). The latter criticisms are particularly troubling because they ignore what makes the theory unique, i.e., it is not the individual parts (or sub-systems) that are so important but rather the way the parts work together — are interconnected — that really matters.

This is not to say that we do not care about the quality of the individual sub-systems — for example, creating high-quality curriculum and instructional methods — but that within the theory of systemic reform is embedded the idea that it is the extent to which the parts work in concert that drives higher system performance. This is where the "cybernetic" perspective comes into play by accentuating the interactions — links and loops — among the sub-systems, particularly the types of information flows that occur (the communications) and the types of control processes that are put in place (the

procedures for obtaining representations of actions and making needed adjustments to get closer to desired goals). It is the complex inter-play among the different aspects of the education "system" that is at the heart of the theory of systemic reform.

Implications for Evaluation

What does this all mean for the development of a plan for monitoring and evaluating the implementation of systemic standards-based reform? First, because we expect a complex system such as education to evolve and change over time, we need to understand the starting point (i.e., the **baseline** state of the policy and governance system) of the system, or what Clune (forthcoming) calls the "prior policy" state, before the introduction of systemic reform. Of course, because systemic reform has been underway in some states for at least a decade and because it has become so pervasive, this may be very difficult to define.

Second, as an evolving system one cannot think of an evaluation as a "snapshot" of how it looks at a particular point in time, but rather what is needed is a "moving picture" that captures a continuous causal chain with feedback loops and ongoing adaptations. This then argues for an ongoing, and long-term, monitoring process rather than a one-shot evaluation if we are to truly understand this new approach to educational improvement, especially since many of the interesting outcomes may take many years to unfold. As Clune (forthcoming) puts it:

> *This type of system is dynamic even in its fully mature state (requiring constant communication and adaptation), and even successful reform will likely proceed incrementally (with more reform leading to gradually stronger policies; leading to gradually stronger curriculum for more students and gains in student achievement), so that systemic reform obviously should be represented as a continuous causal sequence….*

Third, in this same vein, one has to view the theory of systemic change not as a "roadmap" that prescribes one pathway to achieving improved student performance through the types of changes discussed by Smith & O'Day. Instead, a better analogy is the way living organisms (complex systems) evolve over time in response to their local

environment, initial conditions, etc. For example, even though birds are related and have many similarities, there are hundreds of different species with substantially different characteristics that have evolved to allow them to prosper within their unique environments. In the same way, evaluating systemic reform is not about trying to "line up" states or districts against some fixed model of what it should look like, but rather it is about capturing the different strategies that are tested, some of which will fall by the wayside in the "competition" to find the best way to raise system performance. For example, the theory is not strong about whether all the components are needed (e.g., are strong assessments and accountability sufficient?), or the best sequence with which to implement reform (i.e., is it better to start with standards, or as in the case of Maryland, to start with the assessments?).

Fourth, the primary focus of a study of the implementation of systemic reform must be the relationships between and among the parts instead of a detailed study of any particular component. That is, the theory envisions an intervention — systemic reform — that brings about order out of chaos. The result of this order is an aligned and coordinated system in which everyone is headed in the same direction, and the tools for getting there support this goal through alignment and ongoing communication that allows for adaptation and adjustment. It is, therefore, alignment that is the key policy instrument that needs to be tested in an evaluation of systemic reform including examination of factors that support or hinder the creation of such coherence, e.g., lags between policy signals and responses, slippage between policy and practice, etc.

The Rest of This Report

The next two chapters of this report provide additional background for thinking about an appropriate evaluations strategy. Chapter II examines what we know about how the theory has thus far "played out" in the real world including a review of the current status of national implementation (e.g., how many states have standards, how many have assessments, etc.), and the extant literature on what we know about the effects of implementation of systemic standards-based reform. In particular, we look at the features of aligned education systems that have been shown to lead to improved outcomes (e.g.,

professional development grounded in the curriculum to be taught). Chapter III then reviews the available, and limited, literature on the relationship between standards-based systemic school reform and student learning.

The report then moves to the practical considerations of trying to monitor and evaluate the implementation of systemic standards-based reform. Chapter IV begins with a discussion of the objectives of a plan to monitor and evaluate standards-based reform, and then reviews the key characteristics of systemic reform and their implications for a monitoring and evaluation plan. The discussion then turns to consideration of different evaluation design strategies, and concludes with a proposed conceptual research design. Chapter V focuses on how one can measure the implementation of systemic reform, and Chapter VI deals with options for measuring changes in student outcomes.

Chapter II: The Theory in Action

The theory of systemic reform can be summarized by two overarching propositions:

- A system guided by high standards, and assessed and held accountable on the basis of its performance relative to those standards, will perform better than a system without such standards, assessments, and accountability.

- To the extent that standards, assessments, professional development, and curricular materials and other components of the educational system are aligned, the system will function more efficiently.

Despite the seeming simplicity of these statements, the actual implementation of systemic reform in states, districts, and schools is rather complex. The nature of these complexities, and their likely consequences, is important for the development of a meaningful evaluation of systemic reform efforts -- some factors may promote effective implementation and others might work to impede progress. Without an understanding of such factors we could come to incomplete and/or faulty conclusions about the true effects of the systemic reform movement.

The complexities stem from three major sources. First, the very elements of the reform policies differ in consequentially important ways in different state formulations. For example, standards are set in different subjects, and for different grade levels or clusters of grades, and may differ in the extent to which they focus on basic skills or more challenging skills. Similarly, some states conduct assessments for students at nearly all grades and others at just a few. Assessments may also be more or less aligned with a state's standards and passing levels could also be set high or low. States also vary in the amount of “prescriptiveness” attached to their standards and assessments. That is, some states require districts to adopt state standards with little local adaptation or elaboration, while others encourage districts to develop their own set of aligned standards and assessments. In addition, the level of accountability attached to performance involves only modest consequences in some states and serious consequences in others. Reforms also differ in terms of their scope. Some states include a significant level of effort in, for

example, professional development and curricular development, while others pay only limited attention to such supportive activities.

Second, the local context in which the reform operates may influence its implementation and thereby its ultimate impact. In some settings, for example, reform might take hold more easily or more robustly than in other settings.

Third, systemic reform requires a host of decisions and actions spread across different jurisdictional levels including the state, the district, the school, and the classroom. In our loosely-coupled educational system, varying levels of discretion about policy and practice operates at each of these governance levels. As a consequence, different objectives, constraints, incentives, and types of information arise that can affect the shape, and the impact, of the reform effort as it moves from the state to the classroom level. For example, even where state reform policies are highly developed, implementation at the district and school level might be weakened as a consequence of local policies. Conversely, district level policies and practices might lead to high levels of local implementation of systemic reform even in a state with relatively undeveloped state level policies. School and classroom level factors could similarly mediate the intended policy effects.

What We Know

The purpose of this chapter is to review what we know about how systemic reform is working in practice. The review is divided into three sections. The first section describes what we know about variation in state-level reform policies, while the second section reviews evidence on the effects of these policies. The third section focuses on variation in effects as a consequence of local contextual conditions. We should state at the outset, however, that the research evidence on the effects of reform, which we report in the second and third sections, is sparse.

The State of State Policies

States are the primary actors in developing systemic reform policies. A number of groups, including the American Federation of Teachers (AFT), Education Week's *Quality Counts*, the Council of Chief State School Officers (CCSSO), Achieve, Inc., the Council on Basic Education, and the Thomas P. Fordham Foundation, have been tracking the progress of state-level reform efforts. Because the AFT, *Quality Counts*, and Fordham all have reports analyzing progress through the 1998-99 school year, we focus on these three sources for the purposes of this discussion. Their analyses highlight three patterns.

A Major National Systemic Reform Effort is Underway

According to the AFT (1999), a total of 40 states have developed content standards in four core academic subjects (English/language arts, mathematics, science, and social studies), and seven additional states have developed them in at least two of the four subjects. The other groups tracking reform also report high numbers of states with standards. Fordham (2000), for example, reports that 45 states have standards in the five subjects (English language arts/ reading, history, geography, mathematics, and science) identified as "essential" by governors at the 1989 "education summit," and 48 states have standards in at least some of those subjects. Similarly, *Quality Counts* (2000) claims that every state, except Iowa, has standards in at least one subject, and that 44 now have them in the four core subjects, six more states than in 1997-98.

The AFT (1999) also reports that all but three states have, or are developing, assessment systems aligned with their standards in at least one core subject area, typically either reading or math. And *Quality Counts* (2000) claims that 41 states administer tests aligned with their standards in at least one subject, up from 38 in 1997-98; and 21 states have assessments aligned with standards in the four core subject areas, up from 17 in 1997-98. The meaning of alignment, however, is not entirely clear, as discussed below.

But Standards and Assessments Vary in a Number of Important Ways

Clarity, level of detail, rigor, and the number of subject areas and grade levels covered are among the ways in which standards differ across states. For example, only a little more than half (22) of the states that the AFT (1999) identifies as having standards in the four content areas meet the AFT quality criteria for clarity, specificity, and grounding in content (AFT, 1999). Fordham's (2000) quality criteria, which also include specificity and content knowledge, identify only eight states and the District of Columbia as having strong standards across the five subjects that it tracks, though 19 states received high ratings from Fordham for their English standards and 18 for their math standards. The Fordham Foundation's general assessment of the quality of standards is not generous, concluding that overall "…state academic standards were a pretty sad set of norms for the nation's schools and children." (Fordham, 2000).

Importantly, both Fordham and the AFT recognize improvement in the quality of standards. The AFT notes that, in the last year, three additional states met its quality criteria, which brought the total to 22. In 1995, only 13 states were in this category. The Fordham analysis concludes that between 1998 and 1999 the average grade for state standards increased from D-plus to C-minus, and the number of states receiving "honor" grades (A or B) in English rose from 6 to 19, and in math from 12 to 18 during the same period (Fordham, 2000).

Standards in some subjects appear to be in better shape those in other subjects, though there is not agreement among the groups conducting reviews. Both Fordham (2000) and the AFT (1999), for example, rate math standards higher than other standards, but they differ in the relative ratings of standards in other subjects. The AFT (1999), for example, considers science standards as being relatively strong, while Fordham considers science standards the weakest, by far.

State assessments also vary widely across states in a number of dimensions, including the subjects tested, the grades tested, the types of tests, and the degree of alignment with the

state's standards. For example, most states have assessments in place for reading and math, but far fewer states — only 17 in 1998 — also have assessments in science, history, social studies, and writing (Achieve, 1998). State assessments also vary in terms of the types of tests used (e.g., norm-referenced-, criterion-referenced, performance assessment). States are increasingly going beyond multiple choice formats; but, according to *Quality Counts*, only ten states require students to develop portfolios of their written work, or to write extended answers in subjects other than English. States also vary in the extent to which they have developed assessments to accommodate students with special needs.

The actual alignment of standards and assessments is a substantive concern in the systemic reform movement and is beginning to get serious attention. Up until recently, reviewers have simply asked state officials if their assessments were aligned with their standards. Efforts have recently begun to independently evaluate the alignment of standards with assessments. Achieve, Inc., for example, is conducting such analyses for several states.[8] In the first two states that volunteered to undergo scrutiny, Achieve identified a mismatch in standards and assessments. In Michigan, they found "…the state's assessment program was substantially more comprehensive and demanding than… the state's standards" (Achieve, 1998). In North Carolina, they found the "…state's academic standards… to be strong and well balanced, but its assessments not as challenging" (Achieve, 1998). Alignment analyses for other states are currently underway.

Evaluations by different groups of the overall quality of systemic reform across the states do not always correspond well to each other. For example, *Quality Counts* and Fordham provide summary ratings taking standards, assessments, and accountability into account.

[8] Achieve's work is based on that of several organizations, including the Learning and Resource Development Center of the University of Pittsburgh, and several experts, including Norman Webb of the University of Wisconsin-Madison. Webb (1999) has developed and tested a system for assessing the alignment of science and mathematics standards and assessments.

Quality Counts gives 11 states a grade of either A or A-[9]; similarly, Fordham gives what it calls "Honor Roll" status to only 5 states.[10] Only one of these states — North Carolina — received an A-range grade from *Quality Counts.* At least part of the reason for the difference is that Fordham has low regard for the standards developed by national professional associations, such as the National Council of Teachers of Mathematics (NCTM), the International Reading Association, and the National Council of Teachers of English, which many states use as models.

Finally, states vary in the degree of control they give districts in the development of standards and assessments. While most states are in the process of completing these elements, some states require districts to establish their own standards and assessments, typically monitoring that activity through a review or approval process (CCSSO, 1999). Some states prescribe all aspects of the assessment system while others prefer a mixed model, in which a state assessment is used in conjunction with locally-developed or selected measures.

Accountability Policies Lag in Development are Unstable and Vary Across States

Many analysts view accountability as the linchpin of systemic reform, yet it appears to be the component that lags other reform elements in terms of development and implementation. For example, *Quality Counts* (2000) notes that "…only a handful of states made progress last year on holding schools accountable for results." In 1999, 36 states reported they were planning to issue report cards on schools, 19 to provide overall school performance ratings, and 16 had established legal authority to take over, close, or reconstitute schools. In 2000, the numbers will be 40, 21 and 18 respectively, indicating not much progress in the last year (*Quality Counts*, 2000).

State accountability policies include both rewards and penalties. According to *Quality Counts* (2000), 13 states now offer rewards to schools for performance outcomes, and 18

[9] These include: New Mexico, Maryland, New York, North Carolina, Oregon, Massachusetts, Virginia, Florida, Oklahoma, Kansas, and Nevada.

states have schoolwide sanctions. During this last year, two states (Oregon and South Carolina) discontinued their school reward programs and one state (California) established such a program. Only 20 states have a program of assistance to low-performing schools as part of their accountability measures. One of the most ground-breaking accountability initiatives was established in Florida this past year; students in chronically low-performing schools will be awarded vouchers that they can take to another public or non-public school.

Accountability measures often run into political resistance, especially if students are negatively affected. Not surprisingly, parents think it is unfair when the rules change and their child is not promoted, or unable to graduate, as planned. Kentucky's reform, one of the earliest and most ambitious, was significantly re-structured after a flood of negative criticism (Elmore, Abelman, & Fuhrman, 1996). Similarly, political outcries in California, another early systemic reformer, also led to the abandonment of the state's curriculum framework and assessment (called "CLAS") in 1994.

As with students, accountability measures that include consequences for teachers are scrutinized closely by unions. It is interesting to note that the discussion of accountability in the AFT report (1999) focused only on accountability policies that have consequences for students, e.g., measures to end social promotion, establish high school exit exams, and increase standards for college admission. Any discussion of accountability for schools (and by extension teachers) was noticeably absent. *Quality Counts* (2000) views states as currently "…adopting a more cautious, incremental approach to implementing school accountability measures."

Fordham (2000) is specifically concerned with states that have tough accountability measures, but weak standards. Such a set-up, Fordham argues, does more harm than good. According to its criteria, Fordham identifies Kentucky and New Mexico as such states. Again, ratings of this type differ across the organizations conducting the ratings.

---

[10] These include: Alabama, California, North Carolina, South Carolina, and Texas.

For example, *Quality Counts* (2000) rates both the standards and the accountability system in New Mexico as the very highest quality in the country.[11]

The Scope of Systemic Reform Policies Varies Across States, as Does the Sequence in Which Different Elements are Introduced

Many analysts and proponents of systemic reform argue that implementation is likely to be more effective if accompanied by professional development, and curriculum and instructional materials, that are closely aligned with the state's standards. Indeed, some might argue that such alignment is a necessary condition for systemic reform to work at all; the success of reforms that call for higher standards in the classroom depends directly on the quality of instruction in the classroom, and both professional development and instructional materials support and guide classroom instruction. However, state investment in these aspects of reform, and the policies designed to support them, vary across the 50 states.

Quality Counts (2000) reports that ten states require and fund induction programs for new teachers; 17 states require that time be set aside for professional development of all teachers; and 34 states provide at least some funds for local professional development programs. The extent to which these efforts are actually aligned with standards is unclear, however. It is interesting to note that four of the 11 states that *Quality Counts* rated in the A-range for standards and accountability received D's for their efforts to improve teacher quality. The measure used to indicate "effort to improve teacher quality" took into account not only state policies to support professional development efforts, but also state policies designed to select high-quality teachers through various measures (i.e., scores on standardized exams such as Praxis or the individual's field of study), to ensure that teachers teach in field, and to establish high-quality pre-service education.

The Fordham Foundation (1999) also evaluated teacher quality across states, using pre-service criteria similar to that used by *Quality Counts.* As with standards, ratings of

[11] *Quality Counts* uses the AFT ratings to report on the quality of state standards.

state's efforts to improve teacher quality vary across organizations. Texas and Florida were the only states that received A's from Fordham for their teacher quality efforts. *Quality Counts* gave both states D's in this category.

The developmental sequence of the different elements of reform — the order in which standards, assessments, accountability, professional development, and curriculum/ instructional materials are introduced — also differs across states. Fuhrman (1994) observed that some states were putting their early efforts in assessments, others into professional development, and others into the development of standards and curriculum guidance aligned with the standards. Although Cohen & Spillane (1994) suggest that curriculum frameworks should be the basis for developing other aligned policies, no research has compared the effectiveness of different patterns of sequencing the various elements of reform or different levels of investment in them.

State Policy Effects

In this section, we review the available evidence on whether state policies related to systemic reform "work" as expected. We are interested in whether state policies, indeed, affect district and school implementation of reform. The question is an important one in education since there is a large literature documenting the failure of central policies to affect local actions. The connection between policy and practice, especially practice in classrooms and schools, is often weak (e.g. McLaughlin, 1998; Cuban, 1993; Elmore, 1995). Thus, the simple question of whether action at one level of the system has a systematic effect at another level is a basic one that will be a critical aspect of any evaluation of systemic school reform.

We are, of course, also interested in the fundamental question of whether state systemic reform policies ultimately affect student performance. Systematic evidence is sparse on both counts, but there are encouraging signs. A recent study, for example, of nine high-performing, high-poverty schools identified the fact that educators aligned instruction to the state's standards and assessments as a key factor affecting the school's success

(Charles A. Dana Center, 1999). We first review available evidence on various links in the implementation process and then move to evidence on student performance.

A 1997 national study of school districts (Hannaway & Kimball, 1997) found evidence that state actions have a significant effect on the implementation of systemic reform at the district level, at least as reported by district officials. A representative sample of districts in states that were identified by national experts as having a formidable systemic reform program in place in 1997 (Maryland, Kentucky, and Oregon)[12] reported significantly higher levels of progress in central dimensions of reform than other districts in the country. These districts reported greater progress in: aligning curricula and instructional materials with standards; developing or adopting assessments linked to standards; linking professional development to standards; and linking school/district accountability to student performance. Corroborating this finding, the districts in early reform states were significantly more likely than other districts to report that the assistance and information they received from their state was "very helpful" to their reform efforts. In short, to the extent that district reports are related to actual progress, the evidence suggests that states can, indeed, be effective reformers of local behavior.

Studies have also examined the effects of particular components of state systemic reform policies. For example, the results of analyses of the National Science Foundation's State Systemic Initiatives — designed to improve math and science education in a state through systemic reform — suggests that professional development strategies are a particularly important component of reform and can work to alter instruction in the classroom, but only **if** the professional development activities are high quality and long-term (Shields, Marsh, & Adelman, 1998). Cohen & Hill (1998) report similar results: teachers in California who had more extensive professional development experiences that were grounded in what students were expected to learn, related to assessments, and extended over time were likely to change their instructional practice in the classroom. After a comprehensive review of evidence on professional development and math and

[12] Both Maryland and Oregon received 'As' for the quality of their standards and accountability and Kentucky received a B+. The standards for Maryland and Oregon also met the AFT quality criteria.

science, Kennedy (1997) also concludes that effective professional development focuses on specific content, e.g., guided by specific content standards, and is long-term. Examining classroom effects of policies designed to affect instruction is a difficult task, however, and poses difficult conceptual (Spillane & Zueli, 1999) as well as methodological challenges (Mayer, 1999). In short, it is easier to affect teacher attitudes about particular instructional practices than actual behavior (Shields, Marsh, & Adelman, 1998), and it is not obvious how to make reliable sense out of short term observations of classroom behavior.

A few studies have also examined the implementation effects of state assessments and accountability policies directly. Some research suggests that use of new state assessments influences teachers' attitudes and instructional practices, at least teachers who have participated in developing and implementing alternative assessments (Koretz, Klein, & McCaffrey, 1994; Aschbacher, 1994; Herman, 1997; Floden, Goertz, & O'Day, 1995). Other studies, e.g., Firestone, Mayrowetz & Fairman (1998) suggest that high-stakes tests generate considerable activity related to the test, but do not affect basic instructional practices in substantive ways. They further suggest that serious professional development activities are necessary to affect teaching practice in any deep way.

As noted earlier, 13 states have rewards for high-performing schools, and 18 states have sanctions for those that are found to be low-performing. Evidence on the effects of rewards, per se, is scarce partly because it is difficult to separate the effects of the rewards from other aspects of the reform program. Clotfelter & Ladd (1966), however, were able to estimate the effect of a program in Dallas that awarded $1000 bonuses to principals and teachers and lesser amounts to other school staff in "winning schools," schools that had the largest academic gains (with controls for student background characteristics) as well as other laudatory performance (e.g., high attendance rates).

While complicated, the results do suggest some positive effect on teaching and learning. The effects of rewards and sanctions no doubt depend on a number of factors, such as the perceived fairness of the program, the type and amount of award, as well as the structure

of the program itself, especially whether there are some schools that are more or less likely to be winners based on student characteristics. What is clear from experience to date is that reward systems, and accountability systems more generally, are both politically and technically complex (Elmore, Abelman, & Fuhrman, 1996).

Some evidence also suggests that systemic reform promotes improved student achievement. The most recent state-level NAEP results (1998), for example, can be compared with state-level scores in 1992 and 1994. Of the seven states that showed significant improvement in reading[13] since 1992 at least two — Maryland and Kentucky — are recognized as early leaders in systemic reform. And Connecticut is recognized for its efforts to improve teacher quality. Grissmer and Flanagan (1998), using NAEP data and combining scores across grade levels and subjects, also showed that North Carolina and Texas made the largest average gains in test scores from 1990 to 1997.[14] Neither of these states, however, showed statistically significant gains in reading between 1992 and 1998, presumably a period during which reforms were heavily underway.

The evaluations of the National Science Foundation's SSI program also suggest some, at least modest, achievement results under certain conditions. The SSI programs that invested most directly in the classroom with intensive efforts in professional development and the development of instructional materials were the most effective (Laguarda, 1998; Zucker, Shields, Adelman, Corcoran, & Goertz, 1998). Also, there was some evidence that these strategies worked better in smaller states where direct communication between the state and the classroom was easier to achieve (Shields, et al., 1998).

Local Context Effects

The connections between levels in the education system — state, district, school, classroom — are loose, as noted earlier, and state reforms are likely to be mediated in some way by local factors. Reforms might be moderated or amplified or they might even

[13] These states are: Colorado, Connecticut, Kentucky, Louisiana, Maryland, Minnesota, and Mississippi.

be altered at the district or the school level to match local understanding of reform, local constraints, or local objectives. In addition, and importantly, districts might establish policies independently that are more or less aligned with state policies.

A growing case study literature provides testimony to district-level variation. Massell, Kirst, & Hoppe (1997), for example, found that some districts in the nine states where they track reforms were pursuing their own systemic reform policies and, in fact, were ahead of state efforts. Districts tended to see state standards as only one of the resources that they could call on as they developed their curriculum and instructional guidance systems.

Similarly, Goertz, Massell, & Chun (1998) studied district level accountability systems in fourteen districts in five states — three states with "strong" accountability systems and two with "weak" accountability systems – and found district-level variation in both. While districts in "strong" accountability states tended to have similar district accountability measures, some districts elaborated state requirements. In "weak" accountability states, some districts developed measures similar to those in "strong" accountability states on their own; and some schools developed school reporting of student assessment results, in spite of the fact that it was not required by the district. Spillane (1996) also found that districts interpreted state reform policies differently, even within the same state.

In another set of case studies, Mitchell & Raphael (1999) observed that districts whose reform policies were identified as "promising" by state officials approached reform in very different ways. Indeed, some researchers argue that state policies actually promote increased policy-making at the district level (Fuhrman & Elmore, 1990; Spillane, 1996). And some level of local adaptation of reform is necessary to meet local needs and conditions. But to the extent that local policies and state directives conflict in the messages they send to teachers, they may impede efforts to develop coherent

[14] See Chapter 3 for further discussion of systemic reform and student achievement, especially comments on these conclusions from NAEP.

instructional practices at the classroom level (Grant, Peterson, & Shojgreen-Downer, 1996).

Evidence is also beginning to emerge about the ways in which district- and school-level implementation of systemic reform differs. Using a nationally representative sample of school districts, for example, Hannaway & Kimball (forthcoming) found that larger districts reported significantly greater progress in the implementation of elements of systemic reform than smaller districts. They offer two explanations for this finding. First, the larger central office staff and specialized personnel of large districts may offer advantages of in-house expertise and other capacity during a time of change, even if large size is not advantageous for productivity during a time of steady-state. Small districts appear to be at a distinct disadvantage, often reporting not only low levels of implementation of reform, but also low levels of understanding of the elements of reform.

Second, and this is related to the first reason, larger districts report being more intensively involved in information and technical assistance networks than smaller districts, which may facilitate district learning about reform and the implementation of reform policies. In fact, Spillane & Thompson (1997) argue that the quality of district-level implementation is dependent on the district's "…ability to learn from external policy and professional sources." Smaller districts unfortunately appear to be left to sort out the various demands of reform on their own with little district expertise as well as limited outside sources of assistance. Combining the district-level data with data collected from a nationally representative of schools, Hannaway & Kimball (forthcoming) also showed that patterns associated with district size hold at the school level as well: schools in larger districts report greater reform progress than schools in smaller districts.

A recent evaluation of the Eisenhower Professional Development Program (Garet, et al., 1999) had similar findings. Larger districts appeared to have capacities not evident in smaller districts. In particular, they were more likely both to align their professional development with standards and assessments and to provide higher quality professional development than smaller districts. Similar to Hannaway & Kimball (forthcoming), the

authors suspect it is because larger districts have a better developed infrastructure to support reform efforts.

Hannaway & Kimball (forthcoming) and Garet, et al. (1999) both also report discouraging results for high-poverty settings. Hannaway & Kimball found that reform progress in high-poverty districts lagged that of low-poverty districts, taking size into account in the analysis. And, Garet et al. (1999) found that even though high-poverty districts receive more Eisenhower funding, teachers from high-poverty schools are only slightly more likely to participate in Eisenhower professional development activities. An important task for any evaluation should, therefore, be to uncover factors that promote effective reform in high-poverty settings.

Implications for Evaluation Design

The above review suggests that any evaluation of systemic reform will be necessarily complex. Differences in state-level reform policies, as well as variation in the development and implementation of reform policies at the district- and school-level, should be taken into account in the design to obtain a realistic picture of reform policies and their effects. Contextual factors that mediate or enhance differences in state or district reform effort, such as state or district size and reform history, also should be addressed in an evaluation.

In addition, while it is important to measure differences between states in their individual policy elements, systemic reform is differentiated from preceding education movements by the alignment of all of the policy instruments. As stated in the second proposition of systemic reform, a well-aligned system should operate more effectively than a non-aligned system. An evaluation therefore must address variation in both the reform components, and in how the components work together as a system, to produce a complete picture of state- and district-level systemic reform.

Chapter III: Effects on Student Learning

Unlike the work done on the *implementation* of systemic reform, few researchers have examined the effects of this model of school reform on student achievement. In fact, only four studies have attempted to link multiple components of a state's reform efforts with student performance on academic assessments — Grissmer & Flanagan's (1998) examination of gains in NAEP scores in Texas and North Carolina, the evaluation of the NSF State Systemic Initiatives (SSI) by Laguarda, (1998) and Shields, Marsh, & Adelman (1998), Kahle's (1999) closer examination of equity issues in the Ohio SSI, and the results of the USED-sponsored Longitudinal Evaluation of School Change and Performance in Title I Schools reported by Turnbull, et al. (1999). From case studies to surveys, each study approaches the task differently and covers a different time period. These studies also used a variety of methods to measure the nature of the reform policies in place, and different achievement tests to measure student performance. To further complicate the comparisons, the student assessments used in these studies are not necessarily aligned with the respective state's standards, and therefore at risk of not accurately capturing the effects of reform efforts. Other studies have looked at specific reform components (such as a privately-run professional development program, e.g., Merck Institute, 1998, 1999), but these four evaluations are unique for having tried to capture the many facets of reform and their effects on student learning.

Evaluating the effect of systemic reform on student outcomes is difficult for several reasons. The first issue is deciding what to measure — that is, should student achievement simply be conceived of as performance on an assessment or in the broader sense of a decrease in drop-out rates or an increase in participation rates in certain classes? Second, reform efforts may focus on one subject area, such as reading/language arts, or on several academic areas. Determining which subjects and content areas may be most reflective of reform is critical to having a more accurate assessment of systemic reform. Also, a state or district's goals might call for all students to be performing at a certain level or to make specific gains from the previous year. As a consequence, an

assessment of changes in student achievement must be linked to initial starting conditions. That is, are schools that have consistently been high performers expected to make the same kind of increases that are desired for lower performing schools? Finally, many types of assessments (norm-referenced tests, performance-based tests, etc.) may be used to measure student performance. The challenges involved in measuring reform effects on students are discussed in greater detail in Chapter VI.

Four Studies of Reform and Student Achievement

With these challenges in mind, the remainder of this chapter is a presentation of findings from four studies that have attempted to address the effect of systemic reform on student educational outcomes. Overall, it appears that ***student performance made small, but statistically significant gains attributed to state level systemic reform policies***. Though more research is clearly needed, each study provides a glimpse into the relationship between state level policies and student achievement. We first describe each study, including the methodology used and the limitations the researchers themselves indicated were present in their study. Then, we further discuss the difficulty of capturing the effects of reform measures on student performance as revealed in these studies.

Texas and North Carolina's Gains on the NAEP

Grissmer & Flanagan (1998) conclude that the large gains made by Texas and North Carolina on the National Assessment of Educational Progress from 1990 to 1997 were due to the systemic reform efforts that were implemented in those states. These results were determined by averaging the gains across 8^{th} grade math and 4^{th} grade math and reading tests. North Carolina has shown the largest gains in student achievement in math and reading of any state in the nation, and now ranks well above the national average on the 4^{th} grade assessments even though it was ranked near the bottom in 1990.

The authors began their analysis by first trying to rule out rival hypotheses for why these two states could have experienced such large increases in NAEP scores. To do this they examined four variables that are traditionally thought to explain gains in student

performance: real per pupil spending, the ratio of pupils to teachers, the percentage of teachers with advanced degrees, and the average experience level of teachers. During the time of this study, both states were close to, or below, the national average on each of these characteristics, and experienced small changes in these characteristics over time. However, the two states moved in opposite directions with different magnitudes in many of these characteristics and, more importantly, rigorous statistical analyses were not performed to examine the relationship between any of these characteristics and student performance. Despite this, the researchers concluded that "…the large score gains cannot be explained by changes in student or teacher characteristics or spending levels."

In comparative case studies of the two states, Grissmer & Flanagan point to certain similarities between the reform strategies undertaken in both Texas and North Carolina. Both states have established state standards, have developed assessments linked to those standards, created accountability systems tied to their standards, created feedback systems reporting on student performance at the student, classroom, school and district levels, and have expectations for all students to meet the standards. The states also gave teachers and schools more local control and increased flexibility for deciding how to achieve the standards, sustained the assessment and accountability systems, and focused resources on schools with higher numbers of disadvantaged students. The business community played a substantial role in supporting reform movements in both states, sustaining reform while political leadership in each state changed. Through this reliance on public and private sectors, both states developed an infrastructure to support continuous improvement in reform. The researchers conclude "…that the most plausible explanation for the test score gains are found in the policy environment established in each state." They further claim that this policy environment encouraged teachers to change their classroom strategies, leading to higher student achievement.

As Linda Darling-Hammond (2000) points out, however, the authors are on shaky grounds with this conclusion because, for example, in North Carolina "….the new standards and assessments were not on-line until 1995, and the rewards and sanctions

component of the accountability system[15] was not enacted until 1997, so it was clearly not a factor in these trends." Grissmer & Flanagan do admit that "….while there is much anecdotal evidence, there are little solid data that record these changes" in teaching and its relation to reform or to student learning. They further acknowledge that additional case studies of state policies and student achievement are needed to strengthen their argument.

The NSF-sponsored State Systemic Initiatives (SSIs)

Seven SSI's were selected for a study of achievement gains reported in the initial SSI evaluations (Laguarda, 1998). The study was limited to the seven SSIs (out of a total of 25 projects) that were most likely to produce solid evidence of student achievement gains.[16] Most of the assessment data cover only one "round" of testing, and several SSI's tested different grade levels each year. Only two states collected several years of data for a consistent group of grade levels. Four of the SSIs used existing state assessment data for their evaluations, and the other three used data from assessments they developed.

Laguarda found that four SSIs (Louisiana, Montana, Ohio, and Puerto Rico) could attribute small, statistically significant increases in student achievement to their reform initiatives. Students in classrooms taught by SSI-trained teachers outperformed their non-SSI counterparts by between one and eight percentage points. However, Laguarda raised many concerns about the limitations of these findings:

- ***The amount of data is limited***: Most states could not show a pattern of increased student achievement over more than one year of testing, and in two states, the results were based on very small sample sizes.

- ***Uneven evidence of gains***: Gains in student achievement were not made across grades and were inconsistent across years of testing.

---

[15] North Carolina's accountability system includes the state assessments. The assessments for grades K-8 were implemented in 1996-97; the high school assessments were implemented the following year. http://www.dpi.state.nc.us/

[16] The seven SSIs were located in Kentucky, Louisiana, Montana, New Mexico, Ohio, Puerto Rico, and Vermont.

- ***Size of effect were small***: Only one or two SSIs (of the seven) could show definitive positive gains using one-third of a standard deviation as a measure of statistically significant improvement.

Laguarda noted that these limitations were due to the real-world compromises educators are forced to make. Assessment systems available for the SSI initiatives were themselves limited, making it difficult to generate evidence of achievement. For example, Kentucky's state assessment system produces only school-level scores, which prohibits classroom-level analysis.

The SSI's also confronted challenges inherent in trying to link student achievement to SSI activities, including: identifying the correct treatment group and choosing an appropriate comparison group; sustaining comparisons over time; weighing sample size against the cost of administering independent tests; and, the ability to disaggregate the data to explore effects for different types of students. Furthermore, two SSI's were in states with assessments that were poorly aligned to the curriculum and instruction advanced by the SSI effort.

Of the seven SSI's studied by Laguarda, those with the greatest impact on student achievement: (a) provided intensive professional development (at least six weeks in three of the four most successful SSIs); and (b) invested heavily in the development of curricular materials and in training teachers to use the materials. Laguarda points out that these SSIs were the most likely to lead to gains in student achievement because they worked "…most directly and most intensively at the classroom level." SSIs that focused primarily on developing state-level policies "found it much more difficult to produce evidence of changes in student achievement that could be attributed directly to the SSI" (Laguarda, 1998).

Laguarda concludes that in the short amount of time during which student outcomes were studied, it is unlikely that the SSIs had a sizeable impact on large numbers of students or in large-sized gains. Furthermore, she suggests that current and future SSI's use multiple

indicators of achievement to capture the special nature of changes in student achievement sought by most of the SSI's — changes that might not be measured with traditional assessments. She also recommends that the SSI's plan strong evaluations, although this is costly, and that NSF consider taking greater responsibility for designing and implementing a student outcome evaluation. NSF's leadership in this area would prevent SSIs from spending large amounts of money developing and administering tests (e.g., Puerto Rico's SSI spent $1.2 million to replicate NAEP testing over two years).

SSI's and Equity

Looking in greater detail at SSI reform measures, Kahle (1999) documents how the reform efforts in the Ohio SSI changed teaching practices, and in turn led to a narrowing of the performance gap. To assess the reform and link the reform measures with student achievement, Kahle and her colleagues surveyed a random sample of teachers in over 100 schools, observed 12 to 16 schools annually, collected student achievement and attitudinal data in each school visited, and conducted case studies in five schools over five years. This multi-tier design had a strong focus on equity.

Kahle reports that those Ohio districts where reform efforts included intensive professional development on standards-based instruction for more than half of teachers in a school seem to have contributed to the narrowing of the performance gap. This finding held true in districts where state and district policy and curriculum were highly aligned, but not where alignment was minimal. Yet the students who had traditionally scored well continued to perform better than students who traditionally were disadvantaged. Kahle also points out that different types of performance measures may yield significantly different results, especially for urban, African American students, illuminating a potential problem when evaluating reform using student achievement data.

In a separate examination of student-level outcomes in reform, four of the seven SSIs reviewed in Laguarda's (1998) study focused their evaluation on equity. These states disaggregated data to examine the achievement of student populations of interest,

including by race, poverty level, public versus private school enrollment. No evidence of these outcomes is provided in the report, however.

Reform in High Poverty Schools

The Longitudinal Evaluation of School Change and Performance (LESCP) in Title I Schools (funded by the US Department of Education) attempts to capture the relationship between student achievement and state and district reform policies (Turnbull, et al., 1999). Although not a nationally representative sample, teachers in 71 high-poverty schools in 18 districts in seven states were surveyed on their familiarity with, and use of, a variety of reform policies. Each state had engaged in some sort of standards-based reform activity, and most districts had policies to support this state activity. Questions in the survey do not differentiate between state- and district-level policies, however. To evaluate student achievement, the Stanford 9 norm-referenced achievement tests were administered to students in eight math and reading content areas.

Approximately half of the teachers surveyed reported being "very familiar" with their state's standards, assessments and curriculum frameworks, and a similar proportion reported that their curriculum reflected these policy instruments to a "great extent."[17] However, this teacher-reported familiarity with the state standards (or curriculum frameworks), and the extent to which these policy instruments are reflected in their school's curriculum, were not found to be necessarily positively related to the academic achievement of their students. While there was a significant *positive* relationship for bottom-quarter students' math scores and the extent that teachers reported content and performance standards were reflected in their curriculum, math teachers' familiarity with content standards and curriculum frameworks was *negatively* related to student achievement for students in the top three-quarters of their class, and showed *no relation* to growth in reading scores for either bottom quarter or top three-quarters students (Turnbull, et al., 1999). The researchers explain that "this lack of association may reflect,

[17] Teachers may be reporting on familiarity with either state or district policy instruments.

in part, the differences between skills measured on the Stanford 9 tests and the skills emphasized in state standards and state assessments…"

Discussion

The four studies described above indicate many problems that researchers face when trying to evaluate student achievement gains and systemic reform. Part of the difficulty in determining whether reform efforts have led to student achievement gains may come from areas of concern mentioned above: which types of student outcomes should be included (i.e., test scores vs. participation rates in certain classes vs. graduation rates, etc.), and how should those outcomes be measured (using absolute levels of performance versus gain scores, norm-referenced tests versus performance-based tests).

Additional difficulty arises when trying to determine whether state-level or district-level policies had an influence when both levels exist. The LESCP study provides a good example of this difficulty — the questionnaires make no attempt to flush out the state versus district policy instruments in place and their effects on teaching and learning. Also, the non-reform policy characteristics[18] that Grissmer & Flanagan rule out as having an effect on student achievement, may have played a role in the NAEP score gains, but this role was not detected in their study.

The SSI studies indicate that time is critical to evaluating effects on student performance. Kahle's studies took place over five years, and allowed her and her colleagues to detect measurable increases in test scores. The evaluations conducted by SRI International reveal that the effects of reform on student achievement may not be apparent until after collecting many years of data using comparable assessment instruments.

In summary, capturing how gains in student achievement may be related to systemic reform policies has many difficulties. However, since the movement was designed to

counter low student performance, making this link is necessary if policymakers are to determine whether this broad reform strategy if worth its cost. Chapters IV through VI discuss in greater detail how such an evaluation could be conducted.

[18] Grissmer and Flanagan (1998) report that class-sizes shrank, per pupil expenditures rose and teacher education increased in North Carolina. Each of these factors may contribute to growth in test scores, but are dismissed in favor of the standards-based reform movement in that state (Darling-Hammond, 2000).

Chapter IV: Conceptual Evaluation Design

Systemic School Reform

Systemic school reform, as noted in earlier chapters, is intended to shift education from a more traditional compliance orientation — a focus on "what schools should do" (e.g., rules governing the length of the school year, hiring certified teachers, and maximum class sizes) — to an outcome-based expectation of what schools "are expected to accomplish" in terms of higher academic performance for ***all*** students. This new paradigm seeks "...improved teaching and learning and high student performance by connecting otherwise fragmented systems" (USED, 1998). Three principles guide this move from fragmentation to greater "system-ness":

- ***Alignment of Curriculum and Instruction.*** Instruction has been fragmented and uncoordinated and there is a need for alignment among what students are expected to know, what material teachers should cover in class, and how students should be assessed to determine the extent to which they (and their teachers and schools) are making progress toward achieving the expected levels of knowledge and skills. According to the US Department of Education, "Students learn best when they, their teachers, administrators, and the community share clear and common expectations for education. States, districts, and schools need to agree on challenging content and performance standards that define what children should know and be able to do" (USED, 1998).

- ***Governance.*** Achieving the desired level of alignment among the different aspects of instruction is best brought about at the state level, where curriculum frameworks, teacher certification requirements, and testing systems are established, and that these state-level policy changes will "flow down" through districts, schools, and classrooms to induce changes in the practice of teaching. Such changes include A…broad parent and community involvement, school organization, coordinated resources -- including educational technology, teacher preparation and professional development, curriculum and instruction, and assessments -- all aligned to agreed upon standards" (USED, 1998).

- ***Accountability and Incentives.*** In turn, the improved instruction, clarity of expectations, and tough accountability will lead to increases in student academic achievement. "Student success stems from concentrating on results. Education systems must be designed to focus and report on progress in meeting the pre-set standards" (USED, 1998).

This chapter explores the development of a plan to both monitor the ongoing progress of implementing systemic reform, and assess the extent to which such policy changes lead to increased student performance.[19] A key assumption underlying this discussion is that there is a need for an agenda of research on this topic and not just a single-shot study. We have assumed, therefore, that there are different levels of information needs ranging from ongoing reports about the progress of reform implementation, to highly reliable estimates of its impact on schools, teachers, and ultimately students. Some information can be obtained within a relatively short time period, but the very nature of systemic reform, and a realization of what it takes to raise the aggregate level of student achievement, argues for significant "staying power" to allow these complex reforms to sufficiently take hold and for student progress to be observed.

Evaluation Objectives

The obvious place to begin thinking about any evaluation study is to ask ourselves, "What is the question(s) that we are trying to answer?" This is important not only for thinking about the information we need and how it will be used, but also to determine the choice of an appropriate study design, the comparisons we will want to make, and our interpretation of the study results. In our view, a monitoring and evaluation effort for systemic educational reform should be designed to answer three basic types of questions:

- ***What progress is being made nationally in the implementation of systemic reform, and how does implementation vary by state-, district-, and school- level?*** This first question seeks to determine where and how systemic reform is occurring, how implementation varies both across states, districts, and schools, and how implementation is evolving over time as policymakers gain more experience with this new theory.

- ***Is systemic reform "working?"*** Here we are interested in determining if systemic reform — as it is realized in "natural" settings — leads to district, school, and classroom changes, and if such changes lead to subsequent gains in student achievement (and other measures of school success).

[19] How to measure both aspects — the state of implementation and student performance — are discussed in

- ***Can it work?*** The previous question focuses on trying to determine the effect of systemic reform as it is actually implemented, not an idealized version of what its proponents think it should look like. This question changes the focus to, "Can systemic reform, **when implemented at its best**, have an effect on student achievement and other measures of school success?"

We are also interested in the mechanism or process that produces any observed changes, or if not, we need to know why the expected effects were not found.

Challenges and Constraints

In thinking about an appropriate way to monitor and evaluate systemic reform, one must confront a variety of challenges and constraints that, by necessity, will affect our ability to obtain the most desirable answers to our questions. Some of the more salient issues are discussed below.

What's The Right Unit of Analysis?

The choice of an appropriate unit of analysis poses a difficult decision. Because states are, according to the theory of systemic reform, the primary driver of greater alignment and coherence, should they be the appropriate place to focus our attention? That is, should we examine state systems of education, and how they are being reformed, and try to relate these changes to changes in student educational performance? Or, would districts — or even schools — be a better choice realizing that it is these levels of the system that must undergo fundamental change to have an effect of students? Or, in fact, would students be the "right" unit of analysis since it is at this level that reformers hope to see ultimate improvement? Should particular groups of students (e.g., gender- or race-specific groups) be the focus rather than all students?

There is probably no correct answer to these questions, although practical considerations would likely push one to a choice of lower levels of aggregation (districts and schools) because they are less complex. In our view, however, evaluating systemic reform should be about testing the key components of the underlying theory, i.e., determining the

---

subsequent chapters.

functional relationships among the elements of the system (e.g., between standards and assessments), and seeing if different configurations — especially those that lead to greater alignment — make a difference in terms of what happens in classrooms, and how any observed instructional changes lead to better student outcomes. In this view, states would be the starting point for an evaluation, and the investigation should progressively trace the connections among different system components as they are implemented among the different levels of the organization, from states, through districts, to schools, and eventually to classrooms and students.

What's the Treatment Being Studied?

In the research literature, a "treatment" is what we do to a targeted group (individuals or organizational units such as schools) to produce change in one or more outcomes of interest. For example, if we were interested in raising reading skills (the outcome of interest) using a new reading program (the treatment) for early elementary grade school children (the unit of analysis), we would want to know if, in fact, students were "better" able to read as a consequence of having had this particular form of instruction, *compared to the instruction they would have otherwise received*. The instruction they would have otherwise received is what researchers call the "counterfactual," i.e., the alternative to which the treatment is being compared to determine if there is a program impact (i.e., a change that can be attributed to the treatment).

The treatment represented by systemic reform is not so straightforward and easily observed as the above example of a classroom reading program. Instead, it is a complex mixture of policy, governance, and practice reforms that are intended to dramatically change the landscape of American education. Some of the intended changes are tangible, such as new standards and curriculum for ***all*** students; others are more ethereal, such as changed attitudes and beliefs about the value of striving for continuous improvement both in how schools function and in student academic performance. Further, some of the anticipated changes are expected to occur at the state level (creating new standards) and must permeate the entire education system to have their intended effect; other parts of the

treatment, such as professional development, can have a more easily traced effect on instruction (if, in fact, one exists).

One consequence of this multi-faceted, interactive policy environment is that the success of systemic reform will depend upon the actions of many political actors and levels of government to attain the end goal of better student achievement. As a consequence, because the causal relationships are not well understood, it will be hard to attribute the "treatment" of broad policy change to changes in classroom instruction and student learning. That is, the policy intervention must work through a causal chain of expected events to eventually impact student achievement. For example, teachers will be trained in the new standards, curriculum, and assessments with the expectation (or hope) that they will be able to translate the training into improved classroom instruction, that they will be able to sustain these changes in instructional practice, and that these new practices will make a difference in student achievement.

In addition, because systemic reform is not a single program or intervention, the line of demarcation between the "treatment" and "no treatment" world (the counterfactual) is very fuzzy. Commenting on their evaluation of the State Systemic Initiative (SSI), Zucker, et al. (1998) observed that, "…many states (both states that were and were not funded by the SSI program) have supported activities that can rightfully be viewed as part of the nationwide movement to implement systemic reform."

Finally, numerous policy and program factors create a highly variable treatment including differences in the: definition of "challenging" standards, how curricula are aligned to those standards, how assessments are done to measure progress, and decisions to pursue a host of other "school reform" activities. For example, in a recent study of districts receiving Goals 2000 subgrants, Mitchell & Raphael (1999) found substantial differences in what activities were being supported, including: the process of developing new curricula and assessments; providing seed money to start innovative programs to assist low-achieving schools; conducting staff development on performance-based assessment and instruction (including hiring substitutes to cover for teachers during in-service

training); creating a program of teacher "mentors," purchasing computers and training teachers in their use; and, developing and sustaining extended-day and extended-year programs. Further, the local context varies across districts and among schools (e.g., size, poverty, racial/ethnic composition, level of current student achievement, access to resources, history of school reform, etc.), and these variations can have additional implications for how the program will be realized "on the ground."

What Outcomes Should be Assessed?

In addition to understanding the nature of the treatment or intervention, one must specify the criteria that will be used to judge success. Is it implementation of the various parts of the theory to create an "aligned" system? Or is it getting to the level of changes in classroom instruction? Or, should it be changes in student academic outcomes?

In our view, the systemic changes are clearly important as this is what the theory is about, but the ultimate test of is whether greater system-ness makes a difference for students. As Cohen (1994) so well observed, "Students' academic performance is the domain of chief interest in systemic reform, but it also would be the most difficult to evaluate." This topic is further discussed in Chapter VI.

How Can We Determine If Systemic Reform Leads to Achievement Gains?

In the years since Donald Campbell's seminal works on social experimentation (e.g., Campbell, 1971; Campbell & Stanley, 1966; Cook & Campbell, 1979), funders and policymakers have increasingly tried to assess the extent to which social programs "work," i.e., to determine if the program achieves its desired objectives.

Along with this increased interest in evaluating the consequences of policy or program alternatives has come increased sophistication about the importance of considering evidence obtained from rigorous research designs when making decisions about program funding. That is, decision-makers have been forced to move from reasoning about "implicit" models of a program (i.e., "I just know this program is effective"), to "explicit"

models that require the specification of hypotheses and a rigorous test of those expectations. In fact, in addition to their normal reliance on testimony and anecdotes from program advocates, implementers, and participants, legislators and other program funders now routinely ask whether research and evaluation studies have been done on the program, and, in particular, they ask informed questions about the quality of the research design and the resulting validity of the evidence produced by such studies.

Because of this increasing sophistication, studies based on strong research designs are often accorded greater weight in the policymaking process. As an example, the broad policy impact of the Tennessee class size experiments (Finn & Achilles, 1999) has been attributed to the use of a rigorous experimental study. Given the high stakes nature of systemic reform strategies any attempt to conduct an evaluation should worry a great deal about the validity of any conclusions about the program's impact, i.e., any evaluation must be politically persuasive which means, at a minimum, it must be scientifically defensible.

Experimentation — The "Gold Standard" of Evaluation

A common approach to the evaluation of program effectiveness is to simply follow participants (e.g., 1st-grade students) after they complete a particular intervention and see what happens to them (e.g., "Do the students exhibit higher 'achievement' in reading?"). But, this method will not tell us what would have happened to them if they did ***not*** receive the selected treatment. That is, if improved outcomes are found for the program participants (e.g., the children are reading at a higher level), we do not know if this is a result of the intervention itself, or the result of normal "growth" in ability, or the result of some other influence (e.g., intensive test preparation).

As a consequence, the accepted way to define program ***impact*** is to compare outcomes (e.g., reading test scores) for individuals who receive some service (e.g., a new type of reading instruction) with outcomes that would have been observed had the ***same*** participants ***never*** received the particular intervention. Of course, we cannot actually

observe the non-event of the same students ***not*** receiving the treatment program (the counterfactual) so we want to create a comparison group that can represent what would have happened to participants in the absence of the program.

In the physical sciences, and now in the social sciences, the preferred method for creating such comparison groups is the use of an experimental trial in which program-eligible individuals are randomly assigned either to a treatment (the program) or control (no program) group. Experimental designs (the gold standard of evaluation) are considered the most valid way to make attributions of program impact because it is well established that non-experimental studies will ***not*** yield an unbiased estimate of program impact, even using new statistical techniques to "adjust" for measurable biases (LaLonde, 1986; Orr, 1998). If random assignment is not compromised by either the individuals (for example, control group members obtaining services on their own), or program managers (who might, for example, use personal judgment to decide which individuals do and do not receive services), program participants will not differ in any systematic or unmeasured way from non-participants. More precisely, there will be differences between the two groups, but the expected or average value of these differences is zero. Under this design, a simple comparison of outcomes for the two groups yields an unbiased estimate of the effect of the treatment condition.

What Would an Experiment Look Like, and is it Feasible?

Ignoring practical considerations, let us examine what an ideal experiment would look like if we wanted to evaluate the impact of systemic reform on student achievement. If we begin by considering the treatment as systemic change that occurs at the state level, then an idealized experiment would take a sample of students and randomly assign half of them to states implementing systemic reform, and half to states that do not. We would then follow these two groups of students and collect data on their academic achievement over some defined period of time (presumably long enough for the reforms to have an effect). In such a design, the only difference between these two groups of students would be the group to which they were assigned (systemic reform or not), and because this

assignment was randomly determined (i.e., it is un-correlated with the outcome of interest) any observed difference in their average achievement is an unbiased estimate of the program's impact "X" months after random assignment. In other words, we can, in this example, confidently conclude that state systemic reform "caused" any observed differences in measured student achievement.

It is, of course, highly impractical (if not impossible) to randomly assign students in this manner. Moreover, states are not "clean slates" as systemic reform has become pervasive at the state level as discussed in Chapter II. Because we cannot, therefore, manipulate the intervention at the state level, any practical evaluation will have to treat states as a "natural" experiment in which one measures variation in both the approach to, and extent of, the implementation of systemic reform, and tries to relate this varying "intensity" of treatment to outcomes of interest (this is discussed in greater detail below).

If we cannot randomly assign states, can we randomize at a lower level of aggregation? Such hypothetical experiments could include randomly assigning districts within states (or schools within districts) to either conduct systemic reform or to continue with "business as usual," i.e., the control group. In such an experiment, the state would serve as the contextual regime under which the district (or school) reform takes place. That is, the observed impact within districts, or schools, would be contingent upon the systemic changes that occur at the higher level of aggregation. For example, districts located in states with "stronger" systemic alignment are likely to have at least different outcomes than those located in "weaker" states.

On a more practical level, as with states, we cannot randomly assign students to different districts or schools. And because there is already progress in systemic reform at these levels as well, even if feasible, what would be manipulated in such an experiment is different approaches to, and progress in, systemic reform rather than the policy change itself.

Of course, experiments do not have to be defined in terms of organizational levels but could instead focus on sub-systems, e.g., professional development for teachers. That is, rather than attempting to examine the entire system, it may be possible to assess the effect of a particular part of systemic reform. As an example, consider instructional alignment at the school level. An experiment could consist of randomly assigning schools to different "treatments" defined by the presence or absence of (or intensity of the effort to create) coherence among what is taught (the curriculum), how it is taught (the instructional practice), and teacher skills in implementing that instruction (through professional development). Comparing teacher and student outcomes across the randomly assigned groups would help answer questions about the impact of instructional alignment.

It is important to keep in mind that an experiment of this type would mean giving up broader questions about the entirety of systemic reform. Such a study would still be of value despite warnings from Chubin (1997) and others that evaluating systemic reform "…must entail more than measuring the performance of the systems' components" because the theory asserts that changing the relationships among the components will "…produce greater positive effects in the whole system."

Using Non-experimental Methods

With the exception of thinking about conducting experiments to test the effect of specific components of systemic reform, it would appear that the most feasible approach to evaluation will involve the use of some sort of non-experimental research design.

Over the years, evaluators have devised numerous ways of approximating a true experiment and these can be categorized into three groups: (1) **regression discontinuity**, (2) **control strategies** (the use of matching and statistical controls), and (3) **instrumental variables**. Ultimately, research design is about finding appropriate comparison groups that approximate the experimental design, and that allow inferences to be made about whether a program has the impact we expect. That is, we want to identify or create groups that can be compared to the group receiving the treatment to serve as an adequate

representation of what would have happened to the treatment group had they ***not*** received the intervention under study (the counterfactual). The following is a review of the different non-experimental methods that could be used for this purpose.

Regression Discontinuity

Regression discontinuity is one of the strongest non-experimental research designs, and is competitive with random assignment in terms of internal validity but with lower statistical power to detect effects. In its simplest form, members of the targeted group are assigned to a treatment **solely** on the basis of a cut-off score on some pretest measure. For example, students who score at or below a certain test score would be assigned to receive a particular treatment intervention, and those above the cutoff point would not receive the treatment. Both groups are then administered a post-test at the end of the treatment period — if the program had a positive effect, we should see a "jump" or "discontinuity" in the regression line for the relationship between the pre- and post-test scores right at the point of the selected group assignment cutoff score.

This model does not have direct applicability to an evaluation of systemic school because the treatment is a "saturation" model where students are not selected for the program based on their prior performance level. A similar type of evaluation model that may be more applicable, however, is the **interrupted time series** in which data on a large number of pre-test and post-test measures (i.e., the same outcomes measured consistently over time) are collected, and there is a sharp transition from the "no treatment" to the "treatment" condition, i.e., a particular point in time when systemic reform was implemented. By comparing the time series pattern before and after the start of treatment, one can infer the program's impact from the presence (or absence) of an observable shift in the time series that directly coincides with the known point (or pattern in more complex designs) of transition. Because of the existing history of implementing systemic reform, and its inherent fuzziness as a policy intervention, it is unlikely that one would have the ability to specify the pattern of transition from the non-reform to reform conditions, nor is it likely that required stream of achievement outcome data will be available.

Evaluator Control Strategies

The broadest, and most common, category of non-experimental designs involve one or more ways that the researcher attempts to "control" those factors that are hypothesized to make the treated group different from the control or comparison group. An experimental study controls these same characteristics through the use of random assignment that creates, by its very nature, equivalent treatment and control groups. In non-experimental models researchers are forced to devise other strategies to achieve approximately the same result. These approaches fall into two categories: matching, and statistical models (e.g., regression).

Matching. Probably the most common way to create comparison groups is by matching treatment units to "similar" units that are not subject to the intervention. Among these, the most common approach is what are called ***reflexive designs***, or simple pre-post comparisons for individuals in the group that receive the treatment. Basically, pretest measures are obtained prior to the start of the treatment, and these are then compared to similar measurements taken at some point after the treatment has been delivered (the post-test). The average difference between the two measurements (across all individuals in the "treated" group) is then used to represent the impact of the treatment. That is, because the measurements are taken on the same individuals before and after the treatment, any observed changes are assumed to be due to the effect of the particular intervention under study.

The biggest problem with this design is that any number things can commonly occur between the pre-and post-test, besides the treatment, that can affect the difference in these two measurements. For example, changes in the composition of the student body (e.g., a rapid influx of Hispanic students), the adoption of changes in instruction that are not necessarily tied to systemic reform (e.g., reduced class size), as well as a host of other factors can increase average student test scores apart from systemic reform. By ignoring these other influences we run the risk of mis-attributing their effect to the treatment, making it difficult to rule out rival hypotheses about what may have caused any of the

difference between the pre- and post-test. Was it the treatment, or, was it one of the other changes that occurred that led to the observed difference? As a consequence, this type of research design is considered the weakest non-experimental model.

Another approach that somewhat improves the situation is the use of ***matched comparison groups.*** In this model, the impact of the program represents the **difference** in outcomes between the treatment and control group (where the difference can be either on post-test scores only, or the difference between pre- and post-test scores for the two groups). The idea behind this class of designs is that by incorporating the use of a matched comparison group, the researcher can control for other environmental changes that may have an impact on the study outcomes. A key assumption, of course, is that both groups are equally exposed to these other environmental influences, and the ***only*** thing that differentiates them is that one group received the treatment. Introducing a comparison group to serve as the representation of what would have happened to the treatment group also introduces a new form of potential bias, i.e., initial differences between the groups (at the time of the pretest) that exist despite our attempts at matching. In other words, any design that depends on matching is only as good as our ability to match on the "right" variables so that we can rule out as many threats as possible to the validity of the eventual comparison. (Recall, that the comparison group has to represent what would have happened to the treatment group had they **not** been exposed to the particular intervention.)

The use of matched comparison groups is perhaps the most popular non-experimental research design, especially in education, as it gives the appearance of helping to control for other historical or contextual changes that may affect the difference in outcomes between the pre- and post-test. But, since systemic reform is a "saturation" intervention (i.e., it seeks to affect all students in a state and/or district), choosing a well matched control or comparison group will be a serious challenge. Some options include:

- ***Matched schools***: In districts that choose to focus reform on only some schools it may be possible to select matched comparison schools. Matching schools is, of course, not easy as there are so many factors that can differentiate schools, and the typical number of candidates for matching within a district are relatively few, thereby

making the matching process quite difficult. In addition, at best one can only match on "measurable" factors (e.g., size, ethnic composition) which ignores all the "un-measurable" factors (teacher and student motivation) that make schools different in ways that are highly likely to affect student achievement. Further, schools can be in very different communities so one would also have to provide convincing evidence that the treatment and matched comparison schools were exposed to the same events during the time of the evaluation (i.e., between the pre- and post-test measure(s)).

- ***Matched districts:*** This is a similar but more difficult strategy because of the greater differences between districts (both in characteristics and the environment to which they are exposed), and the potentially smaller set of districts from which to select a matched comparison group (i.e., there are few non-reform districts). As a consequence, the difficulty in achieving "comparability" between districts will make it hard to rule out other plausible explanations for any observed differences in student achievement gains.

- ***Cohort designs***: This design would use the fact that there are natural cohorts in states and/or districts that could serve as potential comparison groups -- for example, one could compare achievement scores for a cohort of 3rd graders before the implementation of systemic reform (e.g., SY 1996-97), to the similarly measured achievement scores of a cohort of 3rd graders in the ***same*** jurisdiction (state or district) after the intervention (e.g., SY 1997-98). This can be a good design strategy if there are no significant changes in student body composition over the intervening time period, but is does require data on the same outcome measure(s) for both groups. It can also be strengthened by using several cohorts, or panels, over multiple years.

- ***Capitalizing on staggered intervention***: Because it is unlikely that systemic reform is consistently implemented everywhere, it may be feasible to capitalize on this fact by comparing early- versus late-"adopting" districts within a state. For example, professional development may be spread over two or more years, as may the introduction of a new curriculum and assessments. For this to work, of course, the process that decides which district is an early versus a late adopter (or a high versus low reform site) must be unrelated to the behavior of the individual district. One can imagine, for example, that a state might decide which districts get trained for reasons that are unrelated to student achievement outcomes. In these types of situations, this can be a useful design strategy.

- ***Comparison to population norms (test norms)***. This is a variation on the comparison group model that uses a defined norm group as the standard against which to measure gains. As such, it would require that we use a norm-referenced assessment.

- ***Comparison to normative gains.*** Another variation is to compare student achievement gains to normative expectations (e.g., we expect 3rd graders to learn the following skills in math by the end of the school year). This model is most aligned with the theory of systemic reform, but it also tells us nothing about what gains would have been observed in the absence of the treatment.

Statistical Control Methods. The use of ***regression techniques*** is another common approach, particularly in saturation programs with few individuals who can serve as controls. This strategy, using data from many observations, tries to capitalize on "natural variation" in treatment to estimate the effect of the program net of other measurable characteristics that might affect the outcome (e.g., demographic characteristics). In effect, natural variation in treatment intensity is considered "random enough" to serve as a substitute for a true experiment, after statistically controlling for known and measurable differences among individuals (or organizational units).

For example, by collecting cross-sectional data from a large sample of students (at either a single time point, or better, at multiple time points) one can use statistical regression methods to "control for" extraneous differences between students (their individual characteristics, and those of their family, school, and community), and determine the relationship between more (or less, including no) treatment and greater (or lessor) gains in student achievement. However, this approach is only as good as our ability to capture the right variables that distinguish the treated group from the control group. (Including a measure of treatment intensity could also be useful for this purpose using the implementation scale discussed in previous chapters.)

Another statistical control mechanism is the use of ***simulation*** which, although it may be the least used evaluation method, may be an appropriate research method to consider in this situation given the likely difficulty of collecting the data needed for the other types of research designs. By using either existing data sets, or newly collected data, it may be possible to conduct simulation experiments to at least bound the magnitude of possible impacts that might be plausibly associated with systemic reform.

A final non-experimental strategy, ***instrumental variables***, can be best understood by considering what makes the experimental model so powerful. The reason we know that observed outcome differences between the treatment and control groups is an ***unbiased*** estimate of the program's impact is that the random assignment to either group is, by

definition, un-correlated with either characteristics of the individuals being assigned (both measurable and un-measurable) or the outcome of interest itself. It is this simple lack of correlation that yields comparable groups for comparison, and which provides the basis for making strong causal inferences about the impact of a particular program.

The idea behind instrumental variables, then, is to find a characteristic that acts like random assignment (which is just a special case of instrumental variables). To be an effective "instrument," such a variable has to: a) be measurable for all individuals, b) be correlated with whether an individual gets services (in the treatment group), and c) have no direct correlation with the outcome measure(s). In other words, like the process of random assignment, the only way the instrument can affect the outcome is by affecting who does or does not get the treatment. For example, consider a situation in which we were interested in the impact of a particular health therapy offered by some health clinics and not others. A good candidate instrument, in such a situation, might be the distance any individual lives from a clinic that provides the particular therapy. Distance is unlikely to be directly related to health outcomes, but probably has a great deal to do with why certain individuals elect to be served at one clinic versus another clinic, and as a consequence, to have access to the special therapy.

Clearly, it is easier to come up with good candidate instruments in some research studies than in others, so the main challenge in this design method is to identify appropriate instrumental variables and empirically and theoretically test the validity of the required assumptions. So, what measures could be used as instruments to evaluate the impact of systemic reform? That is, what characteristic can one measure that would, for example, determine whether one state is pushing ahead on systemic reform, and yet which is unlikely to have a **direct** effect on whether student achievement improves? Some possibilities include: the extent to which the state legislature and/or the governor are becoming increasingly more or less supportive of education; similar measures of teacher union support for the new standards, curriculum, and assessments; the presence of a state management information system providing school level monitoring information; and,

state exit examinations for high school graduation (unlikely to directly affect, for example, elementary grade achievement test scores).

What Should We Do?

There are no easy solutions to this design problem, and any choice will involve difficult trade-offs and compromises. The following is what we believe is an agenda of research that can, when put together, provide answers to most, if not all, of the questions that will confront policy-makers and program administrators during the coming years. But, rather than proposing a single study, we have recommended a number of inter-locking studies that will allow the accumulation of evidence to increase the strength of our understanding about systemic reform. It is, however, a "grand plan" that will require significant investment and the ability to stick with the research effort for the required period of time. At the end of this section, we provide recommendations regarding the timing and priority of the different components of the proposed overall research agenda.

A Proposed Conceptual Design for Monitoring and Evaluation

In consideration of the competing demands and constraints, we suggest a multi-level research design that is depicted in Exhibit IV.1 and described below. As noted above, details on what should be measured, and how, at the different levels of this design is discussed in Chapters V and VI.

Level 1 — State Implementation of Systemic Reform

At the highest level of the proposed design we suggest the creation of an ongoing monitoring system to collect annual data from all 50 states on their progress in implementing systemic reform. These state-level indicators should focus on measures of what we expect states to be responsible for in the effort to create systemic education reform: the nature and “quality” of state standards/curriculum frameworks, alignment to

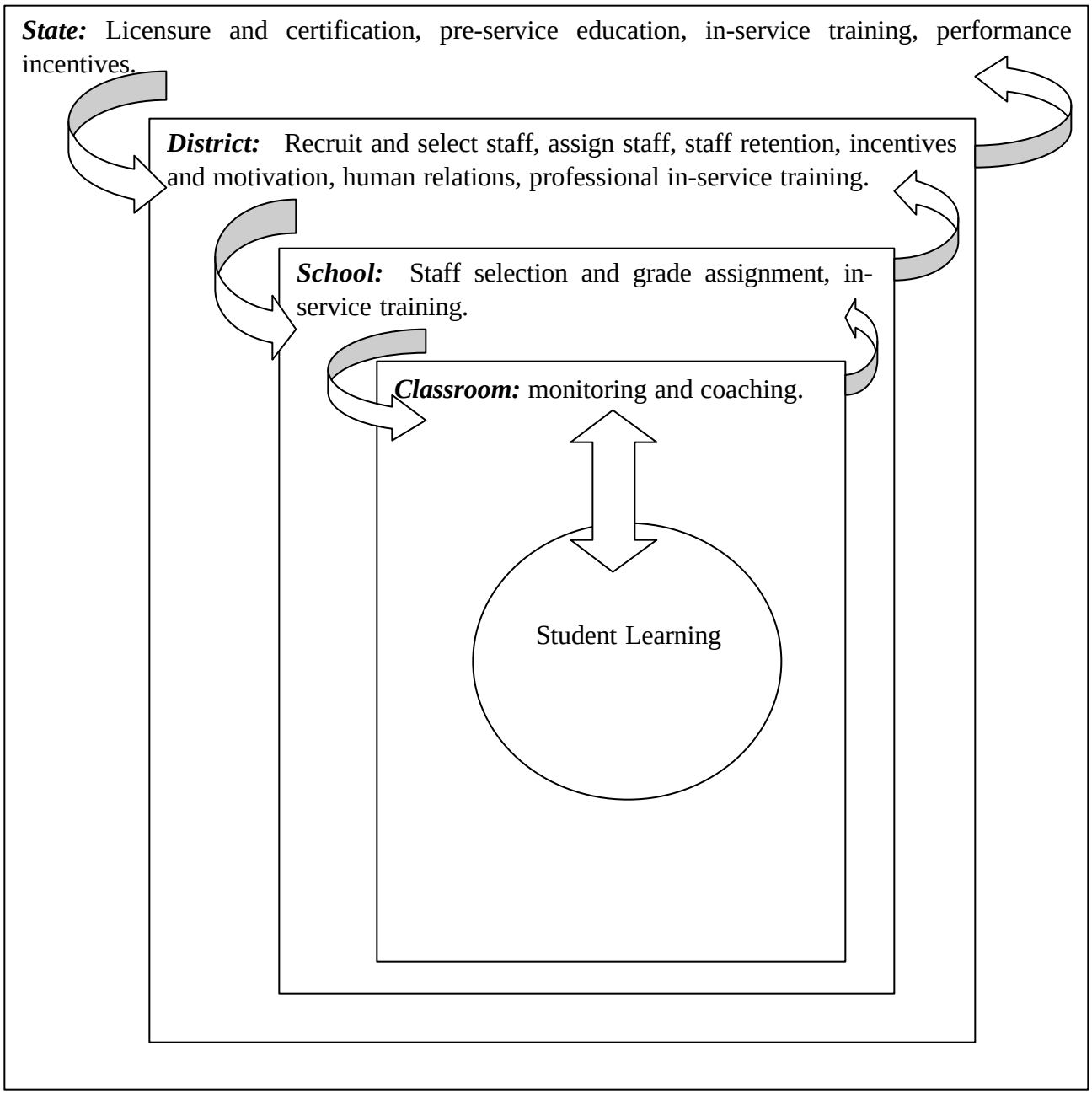


Exhibit I.4: Incorporating Governance Levels: Staff

state assessments, and the strategy they are using to implement systemic reform. It will also be important to capture the key environmental conditions that affect how reform will be carried out, and changed over time, as well as the prior policy conditions that set the "baseline" for measuring policy change over time.

This level of the research plan will answer questions about what states are doing (what responsibilities they have assumed in this reform process), how they are going about it, and the types of contextual factors that either support or hinder implementation.

Currently, the Consortium for Policy Research in Education (CPRE) is developing a framework for describing key components of state systemic reform and accountability systems and collecting data from all states for the framework. Information from this CPRE project could be used to provide data on state reform needed for this Level 1 analysis.

Level 2 — The Relationship Between State Reform and Student Outcomes

The second level of the proposed design is an attempt to find an association between state systemic reform and student progress in academic achievement. This is, of course, a very "gross level" analysis, but it can help shed some light on the extent to which trends in student academic achievement are moving in parallel with the overall policy reforms.

To do this, we suggest using the data collected at Level 1 to create a typology of state-level reform. Because the nature of this categorization will have to be determined by the empirical data, it is hard to define a priori groupings that are likely to emerge from the Level 1 analysis. But, at a conceptual level, one would expect that states will sort themselves into categories that allow some discrimination of different levels of reform "intensity." And, as discussed in Chapter II, early evidence does suggest that this is the case. The framework being developed by CPRE, as well as draft results from the "Pulling in the Same Direction: Goals 2000 and State Standards-based Reform," conducted by

Policy Studies Associates, both collect potentially useful information on state policies designed to support standards-based reform implementation. Appropriate categories might be defined by the quality or "strength" of the standards/curriculum frameworks, and/or the extent of alignment between standards and state assessments, and/or the relative prescriptiveness of the change strategy (i.e., how loose or tight the degree of coupling is among the parts of the state education system).

Once all 50 states have been categorized in this manner, the next step would involve obtaining consistent measures of student achievement that can be compared across states, and that would provide a series of repeated assessments to allow one to judge changes in student performance over time. Studies by the National Research Council demonstrate the infeasibility of two options for the development of a valid measure of student achievement within states and in terms of national performance standards: (1) linking commercial and/or state tests to the National Assessment of Educational Progress (NAEP) and to one another standards (Feuer, et al., Eds., 1998); and (2) embedding test items from NAEP or other tests in state and district assessments (Koretz, et al., 1999). Based on these conclusions, the only available source of such information is the NAEP that serves as the Nation's "report card" on American education (NAEP, 1998). These data can, therefore, be used to determine if there is a relationship between the level (and trend over time) in student achievement and the typology of state systemic reform. In doing this analysis, particular consideration should be given to an overall relationship, and differences in patterns that may exist for particular subgroups (e.g., by race and/or ethnicity). These analyses should be done for all NAEP subject areas for which states have standards, and for the three grade cohorts included in NAEP testing (i.e., grades 4, 8, and 12). Consideration should also be given to conducting similar analyses for other state-level indicators of student progress that can be obtained for most (if not all) states — examples include measures such as average SAT/ACT scores, as well as drop-out and attendance rates.

It is important to keep in mind, however, that this type of analysis — looking for associations between state reform implementation and gains in average NAEP scores or

other measures of student progress — will not provide strong inferences about the "impact" of systemic reform, especially given the different pathways that states are taking to implement this policy. It can, however, provide suggestive evidence of the association between the state's policy reform "model" and gains in student achievement.

Level 3 — District-level Implementation of Systemic Reform

Although interest in state-level reform progress is important, the future viability of systemic reform depends on the ability of states to drive reform down to the level of districts, schools, and eventually classrooms where actual student learning takes place. This is what the remaining levels of the proposed design are intended to evaluate — Levels 3 and 4 focus on districts, and Levels 5, 6, and 7 focus on schools and classrooms.

Level 3 begins with the selection of a national probability sample of school districts to form a "panel survey" in which the same sample of districts is followed over time, with changes in implementation carefully monitored. The selection of this panel sample should involve the use of stratification to ensure the adequate representation of the different state reform implementation models[20] (from Level 1), and different types of school districts (e.g., reflecting characteristics such as urbanicity, and enrollment size[21]).

Data should be collected annually from the panel sample of districts to create a national "snapshot" of district-level implementation of systemic reform. At this level, interest would focus on district alignment with state standards/curriculum frameworks and assessments, and local activities such as the connection between standards and actual curriculum, the development of district-level standards and accountability systems, and local efforts focused on capacity development. As with the state level, information should also be collected on the district's "change strategy," and key environmental characteristics including the prior policy context.

---

[21] Alternatively, one could select districts within stratification cells using "probabilities proportional to size" (PPS) to give the larger districts a greater chance of being selected as they account for the largest proportion of students (i.e., the effect of systemic reform on the average student is more important than the impact on the average district). The trade-off will depend upon USED's desired focus for the study estimates, i.e., the effect for the average student vs. the effect for the average district.

It is also probably worthwhile to augment the information collected from the national panel survey with in-depth case studies for a small sample of districts. This more intensive data collection should be embedded within the panel sample by selecting a probability sub-sample of districts from the larger panel sample and using these sites for the in-depth investigation. Using such probability sampling for both the general sample, and the case study sample, allows the generalization of data to the district, state, and national level.

With support from Planning and Evaluation Service at the U.S. Department of Education, American Institutes for Research (AIR) is presently studying the impact of standards-based reform on instructional alignment and student achievement. The "Moving Standards to the Classroom" research will gather detailed information on district- and school-level policies and practices in six states. Some of this information may be useful to design Level 3 of this analysis, as well as Levels 4 through 6 (below).

Level 4 — The Relationship Between District Reform and Student Outcomes

Using data collected at Level 3, districts can, as with States, be sorted into different categories representing "models" of systemic reform. For example, it may be possible to develop a typology of districts that discriminates on the basis of the "intensity" of their alignment with state policy, or on the basis of their own district-level reform efforts.

Once the districts have been categorized, information on how student achievement has changed over time in the sampled districts will have to be obtained. But, unlike states, there are no currently available, consistent, data on student outcomes available for a large sample of school districts. As a consequence, we recommend that special arrangements be made to have NAEP testing done in the selected panel sample of school districts.

Once these data have been collected, it will be possible to examine the extent to which there is a relationship between the level (and trend over time) in student achievement and district-level implementation of systemic reform (the schools in which testing should be

done are those selected in Level 5, below). Of course, the same caveats noted for the state-level analysis apply regarding making any inferences from this analysis alone.

As above, particular consideration should be given to testing for an overall relationship, and for differences by various student subgroups (e.g., by race/ethnicity). Similar analyses may also be done for other readily available district-level indicators of student progress such as SAT/ACT scores, and drop-out and attendance rates. In addition, additional analysis could be done ***within*** States using scores on State assessments since these districts will have comparable measures.

Level 5 — School-level Implementation of Systemic Reform

A probability sample of schools — stratified by elementary, middle, and high — should be selected from the panel sample of school districts to create an embedded panel sample of schools that will also be tracked over time. The data to be collected at this level should focus on how school-level policy and practice is aligned with state- and district-level policies, and how actual classroom-level practice is aligned with the planned or intended policy.

In addition, we also recommend that schools selected within the case-study sub-sample of districts, be included in the district case studies. This will allow the development of a rich story about how reform implementation occurs within the context of states, districts, and schools. Again, using a probability sample allows the generalization of the survey and case study data to the district, state, and national level.

In addition to information from AIR's study on district- and school-level policies and practices, the National Longitudinal Study of Schools (NLSS), supported by Planning and Evaluation Service at the U.S. Department of Education, is collecting data on the extent to which schools are using standards-based reforms to improve student learning for the 1998-99, 1999-2000, and 2000-01 school years for a nationally representative sample of Title I schools. The successor study to NLSS, the National Study of Title I Schools: Implementation of Standards-Based Reform and Title I Supports for School

Improvement, will collect data from a nationally representative sample of schools, including a nationally representative sample of Title I schools. Data from these two studies could be used to inform Levels 5 and 6 (below) analyses.

Level 6 — Relationship Between School-level Reform and Student Outcomes

Using data collected at Level 5, schools can be categorized in terms of their alignment with overall state and district systemic reform — those that exhibit a high degree of alignment (highly-aligned schools), those who have achieved more moderate alignment (moderately-aligned schools), and those with weak alignment (weakly-aligned schools). As noted above, special arrangements will be made to have NAEP testing done in the selected panel sample of schools to allow for the determination of whether there is a relationship between the level (and trend over time) in student achievement and the extent of the school's alignment with state policy. Also, as above, particular consideration should be given to the determination of an overall relationship, and patterns for specific subgroups (e.g., by race/ethnicity). Similar analyses may also be done for other readily available district-level indicators of student progress such as ACT/SAT scores, and drop-out and attendance rates. In addition, within-State analysis could be done using information on State performance assessments.

As mentioned, data collected from the NLSS study, and its proposed successor study, as well as AIR's "Moving Standards Into the Classroom study, could be tied to Level 6 analysis of the evaluation, as well as Level 7 (below).

Level 7 — Randomized Experiment

At this, the final level, the evaluation should focus on schools found in Level 5 to have "weak" alignment with state systemic reform standards. That is, we want to identify a set of schools that can serve as a "no (or little) treatment" comparison group. To the extent possible, schools should also be selected from districts and states in a way that provides some degree of control over the policy environment. For example, if the data allow it would be a good idea to distinguish a weakly-aligned school in a district that has made great strides toward achieving alignment with the state, from a weakly-aligned school

that is in a district that is less advanced in its reform efforts. Such distinctions would provide some control over differences in district (and, if desired, state) context.

Within the different cells defined by district (and state) reform implementation, weakly-aligned schools would be randomly selected and randomly assigned to either a "treatment" or "control" group. The treatment schools would receive additional resources and technical assistance to implement "high quality" systemic reform to provide a test of whether it can work if implemented well. Baseline data on school characteristics would be collected, as would pretest measures on individual student achievement and a variety of student-level demographic characteristics, ongoing "process" data to measure the level of treatment implementation, and repeated measures of student achievement following the completion of the reform intervention. Comparing differences in student achievement outcomes between the treatment and control groups would provide an unbiased estimate of the activities undertaken to effect high quality systemic reform (controlling for the baseline measures would increase the reliability of the impact estimates).

What would "high quality" systemic reform look like? This is, of course, a critical question because this is the "treatment" that would be tested in the randomized experiment, and providing an answer is made difficult by the fact that the underlying theory is not very specific on this point. Realizing that there are many ways to construct an experiment of this type, the following is one idea that is offered for illustrative purposes only:

- ***Scale of the experiment*** — an experiment of this type will not be cheap to conduct so the desire for information will have to be tempered by the reality of resource constraints. A reasonable scale might, for example, consist of the following:

  1. Select four states, two of which should be implementing "high-quality" standards.

  2. In each state, select four districts, two of which should have strongly aligned curriculum and policies. This yields a total of 16 districts in four states.

     The "Moving Standards into the Classroom Study" being conducted by AIR may provide information for school and district selection.

3. In each of the selected districts, identify weakly-aligned schools. These can also be "failing schools" or schools identified as in need of improvement. Randomly select two of these schools in each district and randomly assign one of them to the treatment group. This yields a total of 32 schools, 16 of which will be designated as the treatment group. (One decision that will have to be made at this point is the school grade level that should be targeted , e.g., elementary, middle, or high schools. For ease of implementation, we recommend concentrating initially on elementary schools.)

- ***Technical assistance provider*** — because this is likely to be a complex effort, each of the selected 16 treatment schools should be paired with a technical assistance provider (e.g., a U.S. Department of Education Comprehensive Assistance Center or Regional Education Lab) to help implement and support the systemic reform process. This partner institution can also help document the implementation of the treatment, and help maintain reasonable fidelity over time. Such providers can include a local university, or state or district staff if the resources and capacity are available.

- ***School year before the experiment is scheduled to begin*** — the prior schools year should be used as a planning and preparation period. Activities to be completed during this time include at least the following:

  1. Work with school management team to review school policies and procedures for alignment with state and district systemic reform goals, standards, and accountability systems. Make necessary changes.

  2. Work with grade-level (or subject area) leaders to review school curriculum, instructional guidance, and materials for alignment with state standards/curriculum frameworks and assessments. To support this process, obtain examples of "high quality" curricula aligned to the specific state's standards from other schools in the district (or from other districts in the state). Develop any needed revisions to the school curriculum and instructional materials.

  3. Prepare for the training of all school-level staff, with primary concentration on the instructional staff.

- ***Conduct Summer Institute*** — because of the effort required, it is probably best to use the summer as a period of preparation for the reform implementation. During this time, all school staff should participate in training designed to motivate and energize them for the coming year, create a new and positive school climate, to improve teaching skills using the new approaches to instruction and assessment, and to review any new school policies and procedures.

- ***Community communication*** — the changes put in place should also be communicated to students, parents, and the community through a variety of media approaches (e.g., meetings, radio/print articles, mailings, etc.).

- ***Ongoing support*** — the technical assistance provider should be available to work with the school leadership, and teachers, on an ongoing basis during the experiment to support the process of systemic school reform. This is in line with the idea that this type of change is not a one-shot effort, but rather a continuous process of improvement.

The schools that are assigned to the control group would receive no additional support and would be allowed to pursue whatever changes they would have attempted on their own — this provides the "counterfactual" against which the treatment will be compared.

Summary

In summary, the proposed multi-level design can:

- Allow for a rich description, and longitudinal tracking, of what states, district, and schools are doing in terms of systemic reform, and how this is evolving in response to both changing conditions and self learning.

- Support an analysis of the relationship between reform implementation (at the state, district, and school level) and the level and gains in student achievement or other measures of school success (e.g., graduation).

- Provide reliable, and internally valid, estimates of the impact of school-level systemic reform on student achievement.

When taken together, these individual streams of knowledge will paint a strong picture of both the state of systemic reform in American schools, and its likely implications for students.

Suggested Priorities and Timing

The plan described above is, admittedly, a large-scale effort. But, without multiple threads of evidence it will be hard to ascertain a clear picture of what the new systemic reform movement has wrought in our Nation's schools. The plan does not, however, all have to be done at once, nor do all the pieces have to be implemented at the same time. In fact, a sequential approach could involve the following:[22]

- ***Implement Levels 1 and 2 early*** — assuming that adequate state-level implementation indicators can be developed, implementing these levels of the design first would provide both national monitoring of broad-based progress in implementation and some linkage to gains in student achievement. Doing this early can also provide Congress with some early information on the effect of the new policy direction.

- ***Implement Level 7 (the randomized experiments) early*** — this component is not strongly dependent on the results of the other Levels, and it can provide early, and highly valid, information on whether systemic reform can improve student learning, when done well. Getting information on this question will probably be more important than understanding the natural variation in implementation and student achievement.

- ***Partially implement Levels 3 and 4*** — select the panel sample of districts and collect a single round of data to allow the sites to be categorized into models of systemic reform. Then conduct the in-depth case studies of the "high alignment" districts to get another picture of what a "strong treatment" looks like and how it may be related to student learning.

Of course, there are multiple ways to configure the timing of the different study components and any final decisions will depend upon the availability of resources and the need for information. The only point to be made here is that we believe that the complexity of systemic reform calls for a multi-pronged, yet coordinated, agenda of research that when combined can help everyone weave together a coherent story both about the challenges inherent in the policy implementation, and the eventual consequences for children. The expectation that a single short-term study will provide the needed information is simply unrealistic.

Finally, it is important to note that systemic reform is not intended to be an event — once done it's over — but a process of change. "The biggest foe of systemic reform is impatience on the part of those who need to see change -- and see it soon" particularly since learning "does not accede to such timetables" (Chubin, 1997). "Ideally, systemic reform should resemble a 'contagion' model: reform spreads and scales up as stakeholders assume ownership of the reforms. 'Going to scale' means changing education systems, by school and district, from the *inside out*. Reform cannot occur as a

[22] What follows is one suggestion for an initial approach to the large-scale evaluation.

result of outside pressure alone, which can build momentum but cannot sustain it.....one does not achieve a state of reform, but engages in a relentless process of reform" (Chubin, 1997).

In a similar manner, any evaluation of systemic reform must be a long-term investment that both observes the unfolding of this process, and allows sufficient time for the intended program impacts to develop and reach a stage where they can be measured. This will require obvious investments of time and resources (probably up to 10 years for complete study), but also the commitment to stay with a research agenda for the long haul.

Chapter V: How Should We Measure Implementation?

This chapter describes key factors in the implementation of systemic reform that are intended to be used in measuring progress made by states and school districts. The selected criteria are based on a review of the existing literature, as summarized in Chapter II, and are based on three guiding principles:

- ***The selected factors will <u>not</u> be used to "rate" implementation.*** Our purpose is to describe the status of the implementation of systemic reform and not to make normative judgements about what states and districts should be doing in terms of educational reform.

- ***Objective measures are more appropriate than self-report measures for a national evaluation.*** Wherever possible, we have tried to select criteria that can be measured on the basis of factual information that evaluators can obtain from policies or other descriptive data, rather than from survey or interview responses that rely on subjective judgments.

- ***The key factors should capture variation in state and district implementation.*** The research literature summarized earlier indicates that considerable variation in implementation exists at the state and district levels. These variations -- in alignment, sequence of development, etc. -- will be useful for testing hypotheses about what configurations of state policy and local context produce positive results in teaching and learning. Some factors already appear to promote the effectiveness of reform efforts (such as intensive, high-quality professional development), while others may inhibit such success.

- ***Our focus is on state- and district-level implementation.*** Because systemic reform theory suggests that the alignment of policies will result in a more efficient, productive education system, we have focused on key state and district policies, and the extent to which policies are aligned. We have ***<u>not</u>*** addressed school-level implementation of systemic reform for two reasons: (1) interest in schools should focus more on how reform is *practiced* rather than systemic changes that seek to bring about changes in classroom instruction; and, (2) a separate study funded by the US Department of Education is evaluating the implementation of standards-based reform in terms of aligned instruction at the school level (American Institutes for Research, 1999).

The key implementation factors are divided into state and district criteria, and are further organized to distinguish the essential components of systemic reform (e.g., standards, assessments, accountability). For each set of factors we have included a rationale, discussing why it is important to capture implementation of this component, a list of individual criteria with a conceptual definition, and a brief indication of potential sources for data collection.

Appendix A details the various criteria used, or recommended, by researchers or organizations for each component of systemic reform. The criteria are listed by component of systemic reform and by source.

State Policies

This discussion begins with the measurement of state-level policies that are related to the implementation of systemic reform.

State Standards

As described in previous chapters, standards and curriculum frameworks are the foundation of systemic school reform, and almost all states have established standards in at least one subject area. Several organizations (e.g., AFT, Fordham, Achieve) have attempted to rate or evaluate state standards, and although they differ on the exact criteria (and methodology) used for their assessments, their ratings can be broken down into two categories: dimensions that address the viability of standards as a policy instrument (i.e., signals intended to modify behavior and practice), and dimensions that attempt to capture the instructional content of the standards (for example, specific authors or literary works, or historical figures or events).

We have chosen to focus only on the policy instrument dimensions for the purposes of thinking about an evaluation of systemic reform. In our view, attempts to rate subject-specific standards for their content have either been too general to be useful, or too prescriptive about what is considered to be appropriate subject area content and instruction. For example, a preference for constructivist teaching would result in very

different ratings of content than ratings favoring an emphasis on "basic skills." Such normative judgements should not be part of a study of systemic reform because the theory does not necessarily prescribe one approach over another. Although we want to know if the standards focus on "basic skills" versus "higher order thinking skills," it is an open empirical question as to which approach leads to greater improvements in teaching and learning.

The following are, therefore, recommended criteria that draw upon the work of AFT, Fordham, McREL, Achieve, CCSSO, and the work of Massell, Kirst, and Hoppe (1997):

- **Types of Standards** — We will want to determine whether the state has developed just content standards (broad statements about what students should know), or both content and performance standards (the specification of what level of performance constitutes mastery of the content standards).[23]

- **Coverage** — This factor captures the subjects in which state content (and performance) standards have been established, and the grade levels or grade clusters used.

- **Focus** — The overall focus of the standards is an important feature of a state's reform effort. In particular, we want to know whether the standards emphasize "basic skills" or higher-order thinking skills such as conceptual understanding, reasoning, and application in real-world situations (McLaughlin & Shepard, 1995). In addition, we should establish whether the standards follow a model in which knowledge is transferred from expert to learner, or a constructivist model, in which knowledge is constructed by learners (and teachers function as coaches).

- **Clarity and Specificity** — We are interested in factors that suggest the likely level of impact that the standards can have on instructional practice in schools. Two factors — clarity and focus — are noted most frequently by the many organizations that rate the quality of state standards. In particular, we want to know whether the standards are *clear* enough for users (including the community, parents, administrators, teachers, and students) to understand the intended learning goals.[24] In addition, we need to establish whether the standards as written are *specific* regarding the knowledge and

[23] Title I requires all states to have performance standards. However, the time of implementation for these standards will vary by state.

[24] For example: the document is written in clear English prose, for the general public as well as for educators; standards describe what is to be taught and learned; the document is clear, complete, and comprehensible to all interested audiences: educators, subject experts, policy makers, and the general public; and, the standards are measurable (i.e., they can lead to observable, comparable results across students and schools) (Finn, C.E., M. Petrilli, & G. Vanourek, 1998).

skills students must learn, and whether they contain enough detail to be converted into school curriculum, and for teachers to develop instructional strategies and lesson plans.[25]

- **Development and Review Process** — The theory of systemic reform, as described by Smith and O'Day (1991), suggests that major policy instruments should be created with the involvement of many stakeholders, and that the timetable should permit extensive review by others, including community members and school staff. Furthermore, the timeline and pace of development (including how long the standards have been in place), as well as regular review of the effectiveness of the standards as policy instruments, are important contextual factors in understanding the potential impact of the standards.

- **Stability** — It is important to assess the stability of the current set of standards (e.g., how long have the same set of standards been in place?), and, if applicable, the monitoring or approval process for these standards.

- **Communication Process** — More than just whether standards exist, we want to know how they have been disseminated to promote their effectiveness as a policy instrument. We also want to know who received the standards and how they were disseminated (e.g., posted on web site, disseminated during professional development seminars), and when districts and teachers received them.

- **Power** — Here we mean the ability of the state to wield force, authority, or substantial influence over those to be affected by the standards, i.e., does the state compel compliance with the standards through the use of legal or regulatory force. For example, Oregon, the only state to *legislate* its standards, requires districts to submit plans for how they will assist students to pass the Certificate of Initial Mastery and Certificate of Advanced Mastery. Districts are required in the legislation to design curriculum and professional development around the standards and provide support for students who do not meet the standards. Alternatively, are the standards simply promulgated as a statement of values (e.g., the "bully pulpit" approach)?

- **Authority** — This criterion seeks to measure the extent to which the individuals (teachers, parents, students) or institutions (districts and schools) that are to be affected by the standards see them as worthy of their allegiance, i.e., do the standards compel acceptance and belief. This can be done either through law, or through appeal to social norms, expert knowledge (e.g., benchmarking, ratings and other activities conducted by outside organizations) that attest to the quality of the standards (Porter, et al., 1988).

[25] For example: the standards must be detailed, explicit, and firmly rooted in the content of the subject area to lead to a common core curriculum (AFT, 1999).

- **Prescriptiveness** — In addition to the power accompanying state standards, we would like to know the extent to which the state dictates what should be done in order to implement the standards (Porter, et al., 1988). This involves the degree of local control in the state (e.g., does the state require or encourage districts to develop their own standards?), policy statements that clarify the expectations of the state (e.g., a statement about the relevance of standards to *all* students), and additional guidance (e.g., curriculum framework) to help teachers and schools use the standards.

Data for most of these criteria can be obtained through a review of existing state documents by "expert panels" (or simple surveys or interviews with state officials). An alternative would be to use current ratings being produced by various organizations, but we recommend against this approach. First, some organizations do not rate all of the states, and as discussed in Chapter II, they use very different criteria for their ratings. Second, to choose any particular rating approach would send a strong, and probably undesirable, federal policy signal.

State Assessments

Assessments, an essential element of standards-based reform, have been adopted by most states in reading and mathematics, and many are in the process of developing tests for other subjects such as science, history, social studies, and writing. The tests developed thus far vary in their approach to assessment (e.g., norm-referenced, criterion-referenced, or performance assessment), the grades and subjects tested, and the degree of alignment with state standards.

The development and use of new assessments has proven challenging for states. In particular, the alignment of state assessments and state standards, key to the state's effort, is technically difficult to achieve and substantiate. Furthermore, the cost and time associated with the use of performance-based items (even when considered well-aligned with the standards) has been demanding enough to encourage states such as Kansas and Arizona to "pull in the reins" on their use. These obstacles suggest important facets in the evolution of this key component of systemic reform.

In November of 1999, the U.S. Department of Education initiated a process for reviewing the establishment of state assessments under Title I of ESEA. The Department's review focuses on evidence that the state systems meet the requirements of the law, including alignment, inclusion, and public reporting, and does not involve direct examination of state assessment instruments. Although the purposes of an evaluation are quite different than those of the Department's review process, the proposed criteria are in keeping with our goal of characterizing the usefulness of state policy instruments for affecting educational practice. These criteria, along with criteria developed by organizations such as AFT will allow us to describe implementation efforts objectively, without making normative judgments. Research by Massell, Kirst, & Hoppe (1997) has also provided key questions on assessment reform that are relevant to this project.

Our recommended set of criteria are as follows:[26]

- **Purpose** — Given the complexity of state assessment systems, which often include multiple measures that accomplish different purposes, we believe that it is helpful to note the intended purpose of each state assessment measure (e.g., assessment of student strengths and weaknesses, program evaluation, improvement of curriculum and instruction, accountability for schools and/or districts, monitor performance of districts, school, individual students, or as a model for local assessment development).

- **Coverage and Type —** This factor identifies the type of test(s) used, and the subject areas and grade levels covered by those tests. With regard to the type of test, we will want to distinguish among: norm-referenced, criterion-referenced, performance assessment, portfolio assessment, etc., as well as the format of test items, i.e., multiple-choice, some performance questions. Finally, various aspects of test administration procedures should be noted since these can have important implications, including the testing situation, scoring procedures, the analysis of the data, and reporting procedures.

- **Inclusion** — Because equity is an area of important concern, we will want to measure the extent to which all students participate in the same assessment process, the use of appropriate accommodations and adaptations (e.g., LEP students are assessed in the appropriate language, tests in other languages are developed as needed), and whether exemptions for particular types of students are reasonable.

[26] In cases where the states require districts to develop their own assessments, most of the factors below apply to the model or criteria specified by the state for the development of local assessments.)

- **Alignment** – Although "alignment" can have many aspects, we are focusing here on alignment of *content*: the extent to which assessments report valid information about the performance of students on content standards at an appropriate level of detail including: comprehensiveness – comparable range of topics covered; emphasis – similar degree of emphasis among topics; similar weighting of "basic" and more advanced skills; consistency with expectations defined in performance standards; and, comparable articulation of coverage across grades and ages.

- **Authority** — The impact of assessment results depends in part on the perception of the quality and relevance of the assessment instruments. As mentioned above, some state reforms have been criticized because educators and/or the public lacked confidence in the tests. States may develop the authority of their assessments by disseminating the results of alignment studies, providing information about how a state test correlates with other well-known tests or other indicators of student performance, providing information about the test development process, and/or sharing the findings of field tests and other studies that corroborate the results of the assessment.

- **Prescriptiveness** — Two aspects of "prescriptiveness" exist. The first is concerned with the balance of state/local control in the assessment system (e.g., state model, in which all students are assessed with a common state instrument; mixed model, in which state assessments are supplemented by state-approved local assessments; or local model, in which state has no common assessment but applies uniform standards to approve and monitor assessment systems developed by each district). The second involves the prescriptiveness of the policy statements accompanying the assessment, i.e., does the state describe the assessment as a model for classroom assessment or for instruction? Or, does the state posit the test as purely a measurement tool?

- **Development and Review Process** — As with standards, the process for developing and reviewing the state assessments is of interest here. In addition, if the state in question encourages the use of local assessments, we want to know how the quality, including alignment, of these assessments is monitored?

- **Stability** — This factor examines how long have the assessments have been in place, and, if applicable, the monitoring or approval process for these assessments.

- **Technical Quality** — This factor captures evidence of the validity and reliability of the state assessment(s). This will not involve a separate assessment of these test characteristics, but a review of existing state information.

Rather than an examination of the assessment instruments — which could only be accomplished by a panel of methodologists — we are recommending the assessment of

the potential effectiveness of the assessments as a policy tool. Such an assessment could be achieved by panels of individuals with experience in developing or using state standards and assessments, who would be asked to review existing state documents (e.g., technical manuals, test design data, blueprints, analyses of scoring data, documents describing field tests, validation studies, test administration instructions, etc.). Where necessary, these data could be augmented with some modest surveys of state-level staff.

State Accountability

Accountability systems are considered by some to be a key driver in systemic reform efforts. Yet, state approaches to implementation of this component have varied considerably, and in many cases have been a politically controversial feature of state reform efforts. For example, some states, such as Virginia, have been criticized when they used new assessments, with unproven reliability, for accountability purposes. Others have been taken to task because nearly all of the tested students "failed" the new test. Little, in fact, is known about which type of accountability system is most effective at affecting teaching and learning, although researchers have noted some dangers to the use of high-stakes testing (e.g., excessive narrowing of the curriculum, teaching to the test, and targeting specific areas of an assessment).

Organizations such as AFT (1999), *Education Week's* "Quality Counts" (1999, 2000) and Fordham (1999, 2000), as well as a variety of individual researchers, have attempted to describe key components of state accountability systems. Here, we focus on key features of accountability systems that assess the "strength" of the policy tool, as conceived by Goertz, Massell, & Chun (1998):

- **Purpose and Audience** — This factor establishes the interrelated purpose and audience of the accountability measures (e.g., reporting progress of schools/districts toward established goals, providing information for policy decisions, promoting students, evaluating teachers, allocating resources). Since we are interested in how the reform components operate as a system, it also is important to know if these purposes are aligned with the other components of state reform such as standards, assessments, local curriculum, and professional development (Smith & O'Day, 1991).

- **Choice of Performance Indicators** — States may use multiple indicators in their accountability systems. In addition to the assessments used, states may incorporate additional measures to gauge performance, such as dropout rates, graduation rates, or course enrollment totals. We need to know which type of indicators are used, and the relative weight given to each type. Also, we need to know if the accountable party is responsible for meeting a set performance level or for achieving a certain gain over the previous year's performance, and if this measure is constant for all held accountable, regardless of their initial performance level.

- **Measures of Proficiency** — States choose their own levels of proficiency on assessments. In order to compare the proportion of proficient students on any given measure across states, this factor looks at how the state defines "proficient performance"; that is, the number of levels and the definition of each level that the state uses to measure student performance. This factor also looks at how the state uses these proficiency levels to determine successful and low-performing schools.

- **Scope** —States may choose to develop assessments for different grade levels all or selected grade levels, and for selected subject areas.

- **Inclusion** — Some students may be excluded from performance measures — it is important, therefore, to know who these students are and what proportion they compose of the total student population.

- **Assignment of Responsibility** — States may hold districts, schools, principals, teachers, or individual students accountable. It is important to know both who is held accountable and how the state determines rewards and sanctions for the accountable parties.

- **Power** — Rewards and sanctions are a key component of high stakes accountability. This factor looks at how successful and/or low-performing schools, districts, educators, or students are recognized (e.g., student performance on assessments, by graduation or attendance rates).

  At the organizational level, we need to know if the state distributes financial rewards or "blue ribbon" status, or some other type of reward for high performance or improvement to districts or schools. It is also important to know if, and what kind of, sanctions the state imposes on low-performing schools and districts (e.g., school closure, takeover, reconstitution, replacement of principals/ teachers, allowing students to enroll elsewhere, loss of accreditation).

  At the individual level, this factor examines the types of rewards and sanctions that exist for teachers, principals, or students (monetary rewards for teachers with classes of high performing students, individual teacher replacement, student promotion, high school graduation requirements, etc.).

Also, at both the organizational and individual level, it is important to know if the state actually carries through with its rewards and sanctions when opportunities to apply them come forth, or if such measures exist only as policy.

- **Stability** — This factor considers how long the accountability system has been in place and any monitoring or evaluation process attached to this system.

- **Information Dissemination** — Public reporting is a low-stakes accountability measure, but an easy way to communicate the results of student performance with the public. There are two aspects to this public reporting factor. The first aspect is how the various components of the accountability system were made public (publication in newspaper, on district web site, state brochures sent to each teacher), and if the system was described clearly so that all affected parties understand how they are being held accountable. The second aspect measures if performance results are made available to the public, the format used (e.g., school/district report cards), and if these results are provided to all stakeholders, including parents, in ways they can easily understand (e.g., in different languages). We also want to know if comparisons between student groups are included in report cards, and which groups are included in the report cards.

- **State Support for Low Performance** — Low-performing schools, districts, or students may need technical assistance for improving performance. This factor, therefore, gathers information on how states identify these low performers and what states do to help them achieve at a higher level. This support may include state education department staff providing direct support and technical assistance on demand, creating a professional development infrastructure to support districts, schools and teachers, brokering information for districts, schools and teachers, and creating or supporting professional networks for teachers, schools, or districts.

- **Development and Review Process** — This factor addresses how and when an accountability system first took effect, and how often it is reviewed and revised (if necessary). These measures shed light on the stability of the system, and the extent to which individuals and organizations will continue to place their trust in the system. This factor will also gather information on the types of precautions the state is taking to avoid "teaching to the test" (e.g., changing test items regularly, secure test administration), and if there is a review process for determining if the accountability system is achieving the purposes for which it was designed.

- **Prescriptiveness** — Finally, this factor examines how much local discretion districts have in state accountability systems. It also indicates if the state encourages districts to develop their own requirements, components, or assessment measures.

State Capacity Building

As mentioned in Chapter II, professional development aligned with the standards is necessary for systemic reform to be effective, and a number of states have begun to include such capacity building in their overall strategy (e.g., some states have pre-service and inservice teacher education plans). In addition, both *Quality Counts* (2000) and Fordham (1999) have developed indicators for state efforts to improve the quality of their teaching staff, and CCSSO (1998) and Massell (1998) have identified state-level strategies to build classroom capacity.

States may undertake many different kinds of capacity-building activities. However, here we seek to collect information only on those activities that are designed to strengthen the state's capacity to implement systemic reform. We therefore focus our criteria of capacity-building efforts that are aligned with the state standards. Examples include: professional development for teachers on how to implement the standards, training on instructional leadership for district superintendents that is focused on systemic reform, etc. The selected criteria are as follows:

- **Purpose** —The purpose may be to help teachers teach to the standards or assessments more effectively, to help teachers learn new curriculum related to the standards, to create a network among teachers, or to develop an infrastructure, among others.

- **Scope** — We want to know what participants do and the focus of capacity-building activities, including involving teachers in development of curriculum frameworks, creating resource banks of curriculum materials, and conducting sessions on instructional leadership for superintendents or principals.

- **Incentives** — The state may reward district or school staff for participating in capacity- building measures, including giving credits toward advanced degrees for participation in capacity building activities, designating districts or schools as flagship models with extra staff, providing additional release time, etc.

- **Prescriptiveness** — This factor examines if professional development is required by the state, how the state supports this requirement (by mandating time or money), the number of days required, and type of professional development required.

- **Target groups** — State capacity building efforts may be focused on pre-service teachers, inservice teachers, or other staff. This factor looks at the type of staff, the grades and the content area toward which these efforts are targeted. Also, state

inclusion of special education teachers and ESL teachers is part of this factor. We also include if there is a state plan that allows teachers who were licensed in another state to practice in the current state, and if and how out-of-state-licensed teachers have to demonstrate knowledge/familiarity with the current state's content standards.

- **Method of Delivery** — In order to have a complete picture of what the state is doing to build capacity to implement reform, this factor addresses the length of time and format used in formal training (one-time workshop, ongoing mentoring or coaching, etc.). It also looks at other types of capacity building (e.g., sabbaticals, resource banks of curriculum materials and other instructional materials, etc.).

- **Development and Review Process** — This factor lists the process involved in establishing the capacity building or professional development plan including: the number of years to develop the first professional development plan; the year that the plan was adopted, and the year it was first implemented; frequency of revision; actors involved in developing the plan (national organizations, content or training experts, staff from other states, teachers, etc). Another important factor is whether there is a mechanism for teachers, schools or districts to indicate their training needs; in particular, the process for review and revision of the assessments (e.g., what data is collected, how is it evaluated, who participates) and the frequency of review. Finally, information on how the state monitors the use of local assessments is included in this factor.

- **Preservice Education** — New teachers may have an opportunity to learn the standards, curriculum and assessments before entering the classroom. This factor captures the alignment between preservice training and a state's systemic reform movement. This can include classes on state standards or assessments, information provided to preservice teachers on state standards, curriculum and assessments, or demonstrations that new teachers understand and can teach state standards.

District Policies

This next section focuses on criteria that seek to describe district-level policies that support (or hinder) the implementation of systemic reform.

District Standards and Curriculum

As discussed in Chapter I, Smith & O'Day (1991) suggest that an important district role in systemic reform is to promote state-level policies while adapting them to local conditions. For example, districts can adopt state standards directly, modify the state standards to address the needs of the local community (i.e., adding standards in technology), or create their own standards based on state standards. Typically, districts

are also responsible for creating detailed curricula and instructional guidance that is aligned with the state-level standards and guidance.

However, case study data suggest that many districts establish and implement reform policies independently of their states, with some evidence that this district activity is greater in states where more reform policies are in place (Hannaway & Kimball, 1997; Goertz, Massell, & Chun, 1998; Mitchell & Raphael, 1999). Understanding such district-level activity, and the state policy context surrounding it, is critical to a reasonable interpretation of implementation efforts. Suggested criteria at this level include the following:

- **Standards** — Districts either *adopt* state standards as written; establish *additional* standards that accompany the state standards; *elaborate* the state standards, providing additional detail (e.g., regarding instructional strands); or *delineate* standards not established by the state, as when the state standards are written for grade spans (grades 3-5) and a district spells out standards for each of those individual grades.

  - **Coverage and Development Process** — We want to know what additional standards, or additional detail, the district has established, including the subject and grade levels covered

  - **Process of Development** — As noted above, districts can adopt state standards or elaborate on these standards. This factor examines the process for a district's adoption/adaption of state standards.

  - **Stability** — As with the state-level components, we want to know how long district standards have been in place and what, if any, processes exist for review of these standards.

  - **Alignment Between District-developed and State Standards** — If the district has elaborated or delineated specific standards, they must be aligned with the state standards — that is, providing detail that enhances the usefulness of the state standards for teachers, administrators, as well as parents and the community, rather than confusing users of the standards. Districts are to provide evidence that their efforts are aligned with the state standards. Here we are concerned with the *elaboration* and *delineation* of state standards only.

  - **Communication Process** — We want to know to whom, how, and when the standards adopted by the district were disseminated.

- **Consequences** — In some cases, districts may attempt to enforce the use of standards in the classroom by attaching consequences to noncompliance. For example, districts may require principals to link teacher evaluation to the review of instructional practice against the standards. To ascertain the use of standards, principals may review lesson plans, conduct classroom observations, etc.

- **Curriculum** — Generally, districts are responsible for establishing clear and specific curricula that are based on state and district standards. States vary in the guidance they provide for curriculum development, but, in systemic reform, districts are responsible for articulating a vision of instruction for all grades.

  - **Coverage and Development Process** — The existence of a district curriculum, including the subject areas and grade levels covered, must be established, as well as the process of development (e.g., mapping previous curriculum to state standards, use of curriculum framework)?

  - **Alignment of District Curriculum** — Detailed information about the process of development ought to indicate whether the district curriculum is aligned with the standards. For example, if teams of teachers in the same content area study the standards, map their current curriculum to those standards, and then establish curricula for standards not already covered, the resulting curriculum will be aligned. Note we are *not establishing the pedagogical appropriateness* of the curriculum; instead, we are focusing on coverage of the same topics and skills, at the same grade levels, as explicated in the curriculum.

  - **Clarity and Specificity of the District Curriculum** — This factor is concerned with the clarity of the curriculum document (i.e., can teachers understand the goals and recommended strategies?), and whether it includes sufficient detail so that teachers can implement the curriculum with confidence that they are meeting the expectations (e.g., are lesson plans and/or sample lessons provided). Also, whether the curriculum is clear enough that other interested audiences can understand it: parents, community members, other educators.

  - **Communication Process** — We want to know how, to whom, and when the curriculum was disseminated, including principals, teachers, and other school staff.

  - **Prescriptiveness** — The extent to which the district details how schools and teachers are to implement the curriculum, as evidenced by the amount and nature of district policy statements and instructional guidance. Also, we want to know whether the district encourages or requires schools to modify or develop curricula (Porter, et al., 1988).

District Assessments

Districts typically administer state tests, and some may adopt or develop additional student tests that can serve a variety of purposes (e.g., measure achievement in between grade levels assessed by the state; assess young children's readiness for school; improve instruction; provide additional information on student achievement to district, parents, teachers, and school; evaluate programs; serve a model for classroom-level assessments). In many cases, these district tests are authentic assessments, which involve student performance aligned with the standards. Sometimes lead teachers and district officials develop rubrics that are used by teachers to score these student performances; this process is intended to help teachers better understand the expectations built into the standards (Mitchell & Raphael, 1999).

We are, therefore, concerned first with district administration of state assessment instruments. Two implementation issues are relevant to that procedure: inclusion and test administration.

- **Inclusion in State Tests** -- Information about the inclusion of *all* students in the state test administration process ought to be collected at the district level as well as the state level to assess whether districts are adhering to this (required) policy. Data regarding the participation of LEP students and students with disabilities in the testing, as well as district policies regarding exemption and accommodations, should be examined.

- **State Test Administration** -- Similarly, we want to learn about *district* level measures to maintain the security of state test administration, including the handling of test booklets and answer sheets and policies regarding makeup tests, etc.

Second, we want to know more about the additional assessments developed or adopted by districts as part of their own assessment systems. Typically, these tests serve as policy instruments in systemic reform because they are linked to the standards, although the power of these tests as policy instruments varies significantly across districts and states. Thus the same factors that are used to describe state assessments (see above) should be used to capture district implementation of district-level assessments. This includes: the **purpose of the tests**, i.e., whether the tests are a required component of the state's accountability system, or an additional measure of student achievement, and if so, what

additional purpose the district test serves; the **specified coverage** of the testing, and the **type of tests used**; the degree of **alignment with state/district standards**; the **development and review process** used, including how long the district assessment has been in place; **district test inclusion policies**; **district test administration procedures**; the level of **authority** that adheres to the tests, including the use of experts to supervise development of the tests; the degree of **prescriptiveness**, such as whether the district encourages the use of school or classroom assessments modeled on the district's assessments; and, the **technical quality** of the tests, including information about scoring, analyzing, and reporting procedures that may shed light on the reliability of the results (such as when districts tests are scored by teachers with varying levels of understanding of new assessment procedures).

District Accountability

In most cases, districts implement an accountability system that is prescribed by the state. But, as discussed in the state accountability section, some states implement a mixed or local model for accountability, in which districts have control over some of the indicators that are used in the state accountability system. In addition, many districts have also established local accountability systems independently of their state systems -- with significant variation in both "strong" and "weak" accountability states (Goertz, Massell, & Chun, 1998). These local systems serve many purposes, including: identification of high- and low-performing schools; allocation of district rewards and sanctions based on these identifications; reporting of school results to the community; and the provision of technical assistance for school improvement (ECS, 1999; NRC, 1999).

As a consequence, we will want to examine any district accountability systems to determine the extent to which it is **aligned with the state accountability system**. This information depends on the following factors: whether the assessments and additional performance indicators (e.g., attendance, dropout data) used in the district system are well-aligned with the state standards; whether the purpose of the district accountability

system is well-matched to the state's goals; and how the district's expectations compare with the state's (e.g., the same or higher expectations).

In addition, some district accountability systems — particularly those used in certain large urban districts — appear to be just as, or even more, powerful a policy tool than the state accountability system. In these districts, the local accountability system has significant rewards and sanctions attached and differs significantly from the state system. In these cases, the following factors, identical to those used to measure implementation of state accountability systems (see above), should be measured: the **scope** of the system; the **assignment of responsibility**; the inclusion of any **rewards and sanctions**; the **development and review process** used; the extent of **prescriptiveness**, including whether schools are encouraged or required to select or develop additional measures for accountability; and, how **information was disseminated**.

District Capacity Building

Spillane & Thompson (1997) argue that the extent of policy implementation by districts depends on their "ability to learn from external policy and professional sources and to help others learn these new ideas," also known as the district's "capacity" for systemic reform. Long acknowledged as pivotal to effective educational reform, district capacity has received increasingly more attention, particularly as researchers have begun to ask the hard question of whether systemic reform policies really do "reach down" into the classroom (Wang, et al., 1993). In fact, district and teacher capacity are viewed by some as one of the greatest challenges to the implementation of reform (Massell, Kirst, & Hoppe, 1997).

In this section, we are concerned with the steps taken by districts to strengthen their capacity to implement systemic reform. Defined broadly, such district capacity-building encompasses the dissemination of products and the provision of training and other

opportunities to help staff implement reform.[27] Professional development, in particular, can greatly support teachers in the delivery of aligned instruction. Although some professional development is offered or required by most states, districts have primary responsibility for assessing the professional development needs of its school staff and providing the training through the use of district and state personnel, outside contractors, and others. Many districts are finding that providing staff development for principals and other building administrators helps them to be better instructional leaders in their school, providing an important boost for lasting change (Mitchell & Raphael, 1999).

For capacity-building to support systemic reform, it must be aligned with state and/or district standards. Research evidence suggests that strong alignment, based on the content of the professional development, is effective in supporting reform. For example, Cohen & Hill (1997) found, in their study of reform in California, that professional development that is embedded in the content and materials teachers are expected to use in the classroom was more effective for teaching and learning than professional development with no such relationship. The following criteria are, therefore, recommended at the district level:

- **Purpose** -- We want to know which reform efforts the various capacity-building activities support (e.g., understanding standards, using new assessment methods, developing a team effort across grades, increasing leadership in the reform effort).

- **Method** – It is also important to monitor capacity-building activities that specifically support district-wide systemic reform. These could include:

  - **Products**, such as lesson plans and models, electronic databases for developing aligned curriculum and assessment activities, and information on instructional strategies. For these products, it will also be important to determine the intended audience.

  - **Professional development activities**, such as workshops, coaching, development of local standards/curriculum/assessments. The participants included (e.g., teachers, principals, aides, special education teachers), the duration of

[27] District capacity, as define by Spillane and Thompson (1997), includes human capital, social capital, and financial resources. We address financial resources in the Contextual Factors section (see below).

professional development, the method of delivery (e.g., one-time workshop, ongoing coaching), and whether the activities are required.

— **Other activities/opportunities** that have been arranged or mandated to support staff in implementing reform (e.g., teacher meetings, common prep periods, participation in networks). Here, too, the participants, duration, format, and degree of prescriptiveness should be noted.

— **Incentives** -- District policies can create incentives for staff participation in capacity-building efforts. These incentives can include: credit for advanced degrees and salary increases based on professional development; stipends; and, release time for teachers who take leadership positions in reform.

Contextual Factors

Each state and district operates in a unique climate and, as discussed in Chapter II, systemic reform is more likely to take hold in some contexts than in others. It is important, therefore, to identify these contextual factors and tease out which outcomes relate to the context for reform and which relate to the policies and practices used to implement reform.

Clune (forthcoming) refers to reform measures that occurred before systemic reform initiatives as "prior reform" or "prior policy." He argues that the history of reform in a state will affect the success of that state's reform measures. Though the "prior policy" variable is not explicitly operationalized in Clune's discussion, we suggest factors that may be indicative of "prior policy." For example, Hannaway & Kimball (forthcoming) found that districts in early reform states reported more progress in establishing the components of standards-based reform than districts in other states. "Early" versus "late" reform states may provide a proxy for the "prior policy" to which Clune refers.

Other researchers suggest a variety of other contextual factors that may affect systemic reform implementation, including the state's political profile, the degree of centralization in the state, or the fiscal climate and other resources. The state may also encourage school choice initiatives over standards-based reform measures, and some states have elected to not participate in the federal programs designed to encourage systemic reform measures.

In addition, student-level characteristics, from the proportion of special education populations to the proportion of LEP students to the overall SES level of the state, district, or school, may affect the influence of standards-based reform measures.

As a consequence, the following criteria are suggested for both states and districts:

- **Reform History** — The extent, nature, and timing of systemic reform measures by states and districts must be noted, including the establishment of higher learning goals (through standards and/or assessments) that preceded passage of the 1994 ESEA legislation. Support for these measures should be monitored, based on specific demonstrations of support rather than subjective judgments by interviewees.

- **Political Climate and Leadership** — It will be important to obtain information about the political climate in the state or district such as the stability of political leaders, important political movements that may bear on education, as well as the length of term of important political and educational leaders (e.g., SEA and LEA officials, principals). We also want to see evidence of political support for the systemic reform movement (e.g., legislation passed, allocation of funds through bonds, stability through change in governors, especially to different political parties) and evidence of support from other sources (e.g., business community, parents).

- **Fiscal Resources** — An important aspect of the state or district's capacity for reform is its finances. Important sources of information include the annual appropriate for K-12 education at the state or district levels, sources for these funds, and other relevant fiscal information such as per-pupil expenditures (adjusted for cost differences), average teacher salaries, and school expenditures.

- **Other School Reforms** — Although this factor is related to the political and fiscal features, the type and scope of school choice initiatives may affect the impact of systemic reform in a state, especially because these factors may occur concurrently with the standards-based reform movement. Information should be collected regarding: school choice options, charter schools, site-based management, flexibility in school regulation, promotion of market-based approaches to education.

- **State/district/school Characteristics** — Size, region of country, poverty rate, staff characteristics (including staff/student ratios, educational background of staff, etc.), policies regarding educational offerings (e.g., AP courses, college credit), overall philosophy regarding centralization.

- **Student Characteristics** — These include data about the student population, class/school statistics, course-taking, home resources (e.g., home computers, parents' educational background).

Data Collection

The key implementation factors described in this chapter are intended to be used to capture state and district progress in establishing and implementing systemic reform policies. In many cases, information on these factors can be collected through a review of state and district documents, including federal program reports from states on performance, policy statements and instruments, plans, records, and dissemination materials. In some cases — such as to document the development of policy instruments or the scope of certain activities — simple surveys or interviews of state and district officials may also be necessary (e.g., superintendent, curriculum and instruction officers, director of assessments and accountability).

However, certain factors related to the nature of the policy instruments will require a different data collection method. These factors include: the focus, clarity, and specificity of standards; the alignment, clarity and specificity of curricula; the alignment and technical quality of assessments; and the alignment of district and state standards.

Many possible methods can be used to assess these factors. One example of how this might be accomplished is to organize three-person panels consisting of individuals who are experienced in developing and/or using standards, curricula, and/or assessments. Ideally, these individuals should come from the ranks of state and local education agencies, as these professionals are especially familiar with the issues raised by the factors. To prepare the panel members, the contractor sends guidance materials to all panel members regarding the key implementation factors named above. Then, a modest training session could be conducted, which can most likely be accomplished long distance. Relevant policy instruments (e.g., standards, assessments, curricula) could be collected from states and districts and distributed to all three members of a panel. In a given time frame, each panel member studies the instruments and makes his/her own determination of quality, alignment, etc., related to the proposed criteria. The panel members then contacts one another by telephone to compare their determinations. If

members disagree, they resolve their differences through a conference call and then submit their final determination.

Analysis Strategy

The measures of reform implementation described in this chapter will produce reasonably extensive data characterizing the status of reform in all 50 states plus the District of Columbia, as well as for a panel sample of school districts as discussed in Chapter IV. To be useful, these data will have to be reduced to a set of summary indicators that can be used to paint a national picture of the status of reform implementation.

For the most part, this will involve the creation of "typologies," or analytical categories, for each of the criteria described in this chapter. We have intentionally avoided an attempt to a priori specify the most appropriate typologies, as it is our view that these are best developed using the actual data that are collected from states and districts, i.e., the creation of a parsimonious set of categories that also provides a good reflection of the existing variation is an empirical, rather than a normative, question. Once the typologies are developed, however, it would be a relatively straightforward process to develop tables that provide the percent of states (or districts) that are estimated to have a particular characteristic (e.g., have performance standards in English/language arts for the 3rd grade). The tables should also provide "breakouts" by key state and district characteristics, as described in the above section on contextual factors. A synthesis of these various tabulations will, then, produce a description of the current status of reform implementation at the national, state, and district level, and show where important variations are occurring.

Another important area of interest, as discussed in Chapter IV, is the need to relate the level or "intensity" of reform implementation to potential gains in student achievement. At one point in this project, the intent was to create an implementation "scale" that would reflect "progress" in systemic reform. However, current evidence suggests that this "single path" model is incorrect as there are probably many routes to reach the same goal of creating a more effective educational system. Second, to be useful, a scale would be

expected to include a requirement that there is a known relationship between the different values of the scale, i.e., that moving across different levels of the scale has a defined meaning. For most of the criteria listed above, this is probably an unrealistic expectation.

As an alternative, we recommend using the descriptive synthesis discussed above (i.e., the tabulation of state and district implementation categories) to create separate variables that can subsequently be used in a statistical model (e.g., linear regression) relating policy and practice characteristics to changes in average student test scores.[28] Because of the natural hierarchical nature of the data it may be best to estimate these relationships using "hierarchical linear modeling" (see, for example, Bryk & Raudenbush, 1992) with the first level of the model capturing the different time point measurements of student outcomes (e.g., the sequence of NAEP tests), the second level capturing state-level characteristics and reform policies, and the third level focusing on district-level policies and characteristics.

Summary

Based on our review of the literature on standards-based reform, we have identified in this chapter the key factors to be used to describe the implementation of standards-based reform at the state and district levels. As mentioned, these factors are organized around the primary components of standards-based reform (e.g., standards, assessments, capacity-building) and, wherever possible, are based on factual information that can be used to describe implementation.

To be used, these factors must be "operationalized" — that is, a measurable indicator must be created for each factor. In some cases, these are obvious (e.g., for how long has the current accountability system been in place?); in other cases, a rubric may need to be developed to measure this dimension (e.g., the prescriptiveness of the district regarding implementation of the curriculum in the schools). We have have attempted to provide details about the specific information that ought to be covered by the indicator. In

[28] These regressions would also control for poverty level, size of school or district, proportion of minority status, etc.

addition, some of the existing and ongoing research on standards-based reform (including work cited in Appendix A and in Chapter 4) can be used as a basis for the creation of indicators.

Chapter VI: How Should We Measure Effects on Students?

In addition to measuring the nature of systemic reform and how it is implemented in states, districts, and schools, an evaluation design will have to include methods for assessing student performance outcomes and how they change over time. This chapter examines this important topic.

What's the Right Construct to Measure?

Student achievement can mean a variety of things to different people. Is it a performance on a single examination or assessment? Is it a broader evaluation of students' ability that covers an extended period of time? Should it, as Hatry (1994) suggests, include other more indirect measures that may be related to attaining higher student achievement, including promotion/retention, attendance, disciplinary actions, and school drop-out? There are no correct answers to these questions, and strong arguments can be made on each side of the issue. In all likelihood, final decisions will be made on the basis of available resources as collecting more indicators of student progress can quickly increase the cost of any evaluation.

Achievement in Which Subject(s)?

Student achievement cannot be assessed in general but must be tied to one or more subject areas. Moreover, reform efforts can vary from place to place — some states/ districts/schools may elect to emphasize the core subjects of reading and math, while others may choose a broader focus on multiple academic areas as part of their systemic reform activities. But, it is both costly and time-consuming to collect information across many areas of academic study. This will require, therefore, that some decision be made about the content areas that will be used as "markers" of program impact. More than likely, these will include the two key core subject areas of reading/language arts and mathematics, and may or may not include other subjects such as science. And, even within a content area decisions will have to be made about what to include as measures of achievement. For example, should writing be included, or is an assessment of reading vocabulary and comprehension sufficient?

What's the Right Standard to Use to Evaluate Performance Gains?

Systemic reform is a process of policy and programmatic change, rather than a single program or intervention. We are interested, therefore, in both the trends and patterns of average performance (i.e., is achievement stable, increasing, decreasing?) and the absolute level of performance (i.e., even if student achievement is increasing, is it below, at, or above a certain proficiency level?).

However, evaluating these results may require a decision about how much gain is "enough" to convince us that the reforms have actually "worked." Setting this threshold for the policy relevance of any observed gains (as opposed to the statistical significance of the difference) is not an easy task. Should we, for example, specify that "Eighty percent of students should master 75 percent of the material tested?" before we would agree that it was a successful intervention? Or, is there a different hurdle that better reflects the intent of the systemic reformers? We could allow policymakers to establish this criterion, but as Ellwein & Glass (1987) found in an examination of state minimum competency tests, not surprisingly, political realities can "soften" the rigidity of a normatively specified cut-score, especially when the results do not support the desires of the policymakers.

A related point deals with whether the same gains ought to be expected of all schools and districts. For example, is it reasonable to expect the same — or even continuously rising — gains for both low-performing schools and for schools that are already at the top of the test-score distribution? On the one hand, we care more about improving scores for the failing schools, and from a psychometric perspective it is easier to make gains at the lower end of the distribution than at the top end of score range.

How Should Student Achievement be Measured?

In order to increase compliance with new state assessment systems, state accountability systems include incentives and support for schools and teachers (Madaus, 1985; Herman, 1997; Linn, 1993). This push for greater accountability has resulted in the nearly universal implementation of statewide student assessment programs. For example, in

academic year 1996-97, 48 states administered some form of statewide student assessment (most often in grades 4, 8, and 11) to improve instruction (47 states), for school-level accountability (40 states), and for student accountability including graduation requirements (25 states); however, only three states reported using assessments as a means of gaining teacher accountability (CCSSO, 1998).

There are, however, large differences in how states have decided to test students, especially as mounting criticism of standardized assessments has pushed states and districts to move away from standardized testing to newer criterion-referenced and/or performance assessments (CCSSO, 1998). For example, 19 states are now administering performance assessments and 4 states use portfolios. Most states include multiple choice questions in their assessments, but 39 states also include written responses.

- ***Type of Assessment:*** writing assessments (39 states), criterion-referenced tests (33 states), norm-referenced tests (31 states), performance assessments (19 states), and portfolios (4 states).

- ***Type of Test Items:*** multiple choice (45 states), extended written responses (39 states), short answers (23 states), examples of student work (10 states), and student projects (4 states).

States that developed criterion-referenced tests and performance assessments are not automatically guarded against criticism about their tests. While new tests may address concerns about standardized tests not reflecting current thinking about human development and may provide teachers with more flexibility in the classroom (instead of "teaching to the test"), issues remain as to the new assessments' reliability, generalizability, fairness, and cost, among other factors. Appendix B discusses these concerns in greater detail.

What Should We Do for An Evaluation?

The tension between the use of traditional standardized testing (e.g., nationally norm-referenced tests) and the newer performance or criterion-referenced tests linked to standards poses a difficulty for the design of an evaluation of systemic reform. The

easiest method of obtaining information on student achievement would be to employ one of the widely used norm-referenced tests. But, using norm-referenced tests would run counter to the intent of systemic reform, which seeks to align standards with curriculum and to align assessments to both the standards and the curriculum. Under this perspective, it would seem to make more sense to use state assessments to better capture what each state/district is trying to accomplish in terms of its own standards and approach to alignment and assessment. However, state assessments are still evolving and are at varying levels of technical quality and quality of implementation. Also, as mentioned in Chapter IV, the use of state tests will prevent comparisons across states since they cannot be equated (although within-state comparisons could be made) (Feuer, et al., Eds., 1998; Koretz, et al., 1999).

As noted in Chapter IV, we have decided to come down on the side of practicality and recommend the use of a single assessment — the National Assessment of Educational Progress — for a national evaluation of systemic reform. In our view, the importance of having a common yardstick to measure performance in a large national sample of states, districts, and schools outweighs the argument in favor of having assessments better aligned to each state's standards. Moreover, because the NAEP tests were developed to align with content standards recommended by national professional organizations (e.g., the National Council of Teachers of Mathematics) they should reflect a close approximation to the standards created in most states. We do, however, suggest the use of state-specific tests that are aligned to standards for special analyses *within* states where the same metric can be applied. Collecting scores for individual students on state-specific assessments, in conjunction with the standard NAEP scores, would also provide a side benefit of permitting an analysis of the relationship between the different types of assessments.

Where possible, we also recommend the collection of other types of student outcome measures to provide a more well rounded picture of the effects of systemic reform. This would include obvious data on the coverage of state- or district-level testing (i.e., who is tested), and other indicators of student outcomes such as SAT/ACT test scores, school

attendance rates, graduation rates, and drop out rates. The latter measure may be particularly enlightening if one of the unintended consequences of systemic reform is to change school drop out rates.

Chapter VII: Bibliography

Abelmann, C., & Elmore, R.F., with Even, J., Kenyon, S., and Marshall, J. (1999). ***When Accountability Knocks, Will Anyone Answer?*** (CPRE Research Report No. RR-042). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Achieve, Inc. (n.d.) ***Putting Education Standards to the Test: Achieve's Benchmarking Initiative.*** Cambridge, MA: Author.

____. (1999). ***1999 National Education Summit Briefing Book.*** Cambridge, MA.: Author.

____. (1998). ***Aiming Higher: 1998 Annual Report***. Cambridge, MA.: Author.

American Federation of Teachers. (1999). ***Making Standards Matter 1999: An Update on State Activity.*** Washington, DC: Author.

American Institutes for Research. (1999). ***Moving Standards to the Classroom: The Impact of Goals 2000 Systemic Reform on Instruction and Achievement***. Draft. December 14, 1999.

Anderson, R.D. (1995). ***Final Technical Research Report: Study of Curriculum Reform. Volume 1: Findings.*** Washington, DC: U.S. Department of Education; Boulder, CO: Studies of Educational Reform.

Asayesh, G. (1993). "Staff development for improving student outcomes." ***Journal of Staff Development***, 14(3).

Aschbacher, P.R. (1994). "Helping Educators to Develop and Use Alternative Assessments: Barriers and Facilitators." ***Educational Policy***, 8, 202-223.

____. (1993). ***Issues in Innovative Assessment for Classroom Practice: Barriers and facilitators***. (CSE Technical Report No. 359). Los Angeles, CA: UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. (1994). "Learning based assessments of history understanding." ***Educational Psychologist***. 29(2), 97-106, 1994.

____. (1994). "Researchers and assessment policy development: A cautionary tale." ***American Journal of Education***, 102(4), 450-478.

Baker, E.L. & R.L. Linn (1997). ***Emerging Educational Standards of Performance in the United States.*** Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Baron, J.B. (1991). "Strategies for the development of effective performance exercises." ***Applied Measurement in Education***, 4(4), 305-318.

Baxter, G.P., R.J. Shavelson, K.A. Brown, & J.R. Valades. (1993). "Mathematics performance assessment: Technical quality and diverse student impact." ***Journal for Research in Mathematics Education,*** 24(3), 190-216.

Bennett, C.I. (1995). "Preparing teachers for cultural diversity and national standards of academic excellence." ***Journal of Teacher Education***, 46(7).

Berman, P. (1982). "Learning to comply." ***Peabody Journal of Education***, 60, pp.53-65.

Berman, P. & M.W. McLaughlin. (1978). ***Federal Programs Supporting Educational Change: Vol. VIII. Implementing and Sustaining Innovations.*** Santa Monica, CA: RAND Corporation.

Beyerbach, B.A., Weber, S., Swift, J.N., & Gooding, C.T. (1996). "A School/ Business/ University Partnership for Professional Development." ***The School Community Journal***, 6(1): 101-112.

Bishop, J. (1998). ***Do Curriculum-Based External Exit Exam Systems Enhance Student Achievement?*** (CPRE Research Report No. RR-040). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Blank, R.K., Pechman, E.M., & Goldstein, D. (1997). ***How Are States Disseminating and Implementing Standards and Frameworks? Strategies in Mathematics and Science.*** Washington, DC: Council of Chief State School Officers.

Blank, R.K., & E.M. Pechman (1995). ***State Curriculum Frameworks in Mathematics and Science: How are They Changing Across the States?*** Washington, DC: Council of Chief State School Officers.

Bruckerhoff, C. (1997) ***Lessons learned in the evaluation of statewide systemic initiatives***. Chaplin, CT: Curriculum Research & Evaluation.

Bryk, A.S. & S.W. Raudenbush (1992). ***Hierarchical Linear Models: Applications and Data Analysis Methods.*** Newbury Park, CA: Sage Publications.

Campbell, D.T. (1984). "Can we be scientific in applied social science?" In R.F. Connor, D.G. Attman, and C. Jackson (Eds.), ***Evaluation Studies Review Annual,*** 9, 26-48. San Francisco, CA.: Sage Publications, Inc.

____. (1979). "Degrees of freedom and the case study." In T.D. Cook & C.S. Reichardt (Eds.)***, Qualitative and Quantitative Methods in Evaluation Research***. Beverly Hills, CA: Sage Publications, Inc.

____. (1971). "Methods for the experimenting society." Paper presented at the annual meeting of the American Psychological Association.

Campbell, D.T., & J.C. Stanley. (1966). ***Experimental and Quasi-experimental Designs for Research***. Chicago, IL: Rand McNally.

Cannell, J.J. (1987). ***Nationally Normed Elementary Achievement Testing of America's Public Schools: How all 50 States Are Above the National Average.*** Daniels, WV: Friends for Education.

Carmine, D. ***Policy Options for Standards-Driven Reform: Standards, Assessment, Accountability, Improvement***. Eugene, OR: National Center to improve the Tools of Educators.

Carter, T.L. (1995). "Continuous Improvement at Skyview Junior High." ***Teaching and Change***, 2(3): 245-62.

Catterall, J.S., Mehrens, W.A., Ryan, J.M., Flores, E.J., & Rubin, P.M. (1998). ***Kentucky Instructional Results Information System: A Technical Review***. Frankfort, KY: Author.

Century, J.R. (1999). ***Evaluators' Roles: Walking the Line Between Judge and Consultant***. Paper Presented at 1999 NISE Fourth Annual Forum, Washington, DC.

Charles A. Dana Center, University of Texas at Austin. (1999). ***Hope for Education: A Study of Nine High-Performing, High-Poverty, Urban Elementary Schools.*** Washington, DC: U.S. Department of Education.

Chubb J.E. & T.M. Moe. (1990). ***Politics, Markets, and America's Schools.*** Washington, DC: Brookings Institution.

Chubin, D.E. (1999). "Findings about systemic reform from evaluations and research." NISE Fourth Annual Forum, Panel Papers.

____. (1997). "Systemic evaluation and evidence of educational reform." In D.M. Bartels & J.O. Sandler (Eds.), ***Implementing Science Education Reform: Are We Making an Impact?***, Washington, DC: American Association for the Advancement of Science.

Chubin, D.E., E.R. Hamilton, & B. Anderson (1999). "Evaluative findings on systemic reform: Lessons from NSF." NISE Fourth Annual Forum, Panel Papers.

Cibulka, J.G. & R.L. Derlin. (1998). "Accountability Policy Adoption to Policy Sustainability: Reforms and Systemic Initiatives in Colorado and Maryland." ***Education and Urban Society***. 30(4): 502-515.

Clotfelter, C.T. & H.F. Ladd. (1996). "Recognizing and Rewarding Success in Public Schools." In Helen F. Ladd (ed.), ***Holding Schools Accountable: Performance-Based Reform in Education.*** Washington, DC: Brookings Institute.

Clune, W.H. (Forthcoming) ***Chapter 2: Building and testing a theory of systemic reform.*** From unpublished book manuscript *Theory of Systemic Reform* by W.H. Clune, E. Osthoff, and P. White.

____. (1999). "Quantitative and qualitative data in the theory of systemic reform." Paper presented at the 1999 Fourth Annual NISE Forum.

____. (1998). ***Toward a Theory of Systemic Reform: The Case of Nine NSF Statewide Systemic Initiatives***. Research monograph number 16. Madison, WI: NISE.

____. (1993). "The best path to systemic educational policy: Standard/Centralized or differentiated/decentralized?" ***Educational Evaluation and Policy Analysis***, 15(3), 233-254.

____. (1991). "Systemic educational policy: A conceptual framework." ***Designing Coherent Education Policy: Improving the System***. San Francisco: Jossey-Bass Publishers.

Clune, W.H., & Witte, J.F. (Eds.) (1990). ***Choice and control in American education, vol.2: The practice of choice, decentralization, and school re-structuring***. New York, NY: Falmer Press.

Coffman, W.E. (1993). "A king over Egypt, which knew not Joseph." ***Educational Measurement: Issues and Practice***, 12(2), 5-8.

Cohen, D.K. (1996). "Standards-Based School Reform: Policy, Practice, and Performance" in ***Holding Schools Accountable: Performance-Based Reform in Education***. Washington, DC: The Brookings Institution.

____. (1994). "Evaluating systemic reform." In US Department of Education, ***Issues and Strategies in Evaluating Systemic Reform.*** Washington, DC: Author.

Cohen, D.K. & H.C. Hill. (1998). ***Instructional Policy and Classroom Performance: The Mathematics Reform in California.*** CPRE Research Report Series, RR-39. Philadelphia: Consortium for Policy Research.

Cohen, D.K. & J.P. Spillane. (1993). "Policy and Practice: The Relations Between Governance and Instruction." In ***Designing coherent education policy: improving the system***. Susan H. Fuhrman. (ed.) San Francisco: Jossey-Bass.

Coleman, J. (1991). ***Parental Involvement in Education***. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement. (ED334028).

Committee for Economic Development. (1985). ***Investing in our children: Business and the Public Schools***. Washington, DC: Author.

Consortium for Policy Research in Education. (1999). ***A Close Look at Effects on Classroom Practice and Student Performance: A Report on the Fifth Year of the Merck Institute for Science Education 1997-98.*** (CPRE Research Report.)

____. (1998). ***Expanding the Breadth and Effects of Reform: A Report on the Fourth Year of the Merck Institute for Science Education 1996-97.*** (CPRE Research Report.)

____. (1993). ***Developing Content Standards: Creating a Process for Change***. (CPRE Policy Briefs.)

____. (1991). ***Putting the Pieces Together: Systemic School Reform*** (CPRE Policy Briefs). New Brunswick: Rutgers, The State University of New Jersey, Eagleton Institute of Politics.

Cook, T.D., & D.T. Campbell (1979). ***Quasi-experimentation: Design and Analysis Issues for Field Settings***. Boston, MA: Houghton Mifflin Company.

Corbett, H.D., & B.L.Wilson (1991). ***Testing, Reform, and Rebellion***. Norwood, NJ: Ablex Publishing.

Corcoran, T. (1995). ***Helping Teachers Teach Well: Transforming Professional Development***. Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Corcoran, T, P. Shields, & A. Zucker (1998) ***The SSIs and professional development for teachers***. Washington, DC: National Science Foundation.

Council for Basic Education. (2000). ***Council for Basic Education Home Page***. http://www.c-b-e.org

Council of Chief State School Officers. (1999). ***Status Report: State Systemic Education Improvements.*** Washington, DC: Author.

____. (1998a). ***Trends in State Student Assessment Programs***. Washington, DC: Author.

____. (1998b). ***Key State Education Policies on K-12 Education: Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance; A 50 State Report*****.** Washington, DC: Author.

____. (1997a). ***Tool Kit: Evaluating the Development and Implementation of Standards***. Washington DC: Author.

____. (1997b). ***Mathematics and Science Standards and Frameworks***. Washington, DC: Author.

____. (1996). ***Measuring Results: Overview of Performance Indicators***. Washington DC: Author.

____. (1991). ***Families in School: State Strategies and Policies to Improve Family Involvement in Education. A Four-State Case Study***. Washington, DC: Author.

Cuban, L. (1995). "The hidden variable: How organizations influence teacher responses to secondary science curriculum." ***Theory Into Practice***, 34: 4-11.

____. (1993). ***How teachers taught: Constancy and change in American classrooms, 1890-1980***. (2nd ed.) New York: Teachers College Press.

Darling-Hammond, L. (2000). "Teacher quality and student achievement: A review of state policy evidence." ***Education Policy Analysis Archives***, 8(1), http://epaa.asu.edu/epaa/v8n1/.

____. (1997). "School Reform at the Crossroads: Confronting the Central Issues of Teaching." ***Educational Policy***, 11(2): 151-66.

____. (1994). "Equity issues in performance based assessment". in M.T. Nettles & A. L. Nettles (Eds.), ***Equity and Excellence in Educational Testing and Assessment***. Boston: Kluwer, pp. 89-114.

____. (1991). "The implications of testing policy for quality and equality." ***Phi Delta Kappan***, 73(3): 220-225

Darling-Hammond, L., & A E. Wise (1985). "Beyond standardization: State Standards and School Improvement." ***The Elementary School Journal***, 85: 315-336.

Darling-Hammond, L. & D.L. Ball. (1998). ***Teaching for High Standards: What Policymakers Need to Know and Be Able to Do.*** CPRE Joint Report Series, co-published with the National Commission on Teaching and America's Future, JRE-04.

____. (1998). ***What can Policymakers Do to Support Teaching to High Standards?*** CPRE Policy Bulletin.

Dauber, S. & Epstein, J. (1993). "Parent Attitudes and Practices of Parent Involvement in Inner-city Elementary and Middle Schools." in Chavkin, N.F. (Ed.). ***Families and Schools in a Pluralistic Society***. New York: State University of New York Press.

Davidson, N., Hinkelman, J. & Stasinowsky, H. (1993). "Findings from a NSDC status survey of staff development and staff developers." ***Journal of Staff Development***, 14(4).

Davies, A. & P. Williams. (1997). "Accountability: Issues, Possibilities, and Guiding Questions for Districtwide Assessment of Student Learning." ***Phi Delta Kappan***, 79(1) pp. 76-79.

Dornbusch, S.M. & Ritter, P.L. (1988). "Parents of High School Students: A Neglected Resource." ***Educational Horizons***. 66: 75-77.

Dorr-Bremme, D., & J. Herman (1986). ***Assessing Student Achievement: A Profile of Classroom Practices.*** Los Angeles, CA: University of California, Center for the Study of Evaluation.

Dunbar, S.B., Koretz, D.M., & H.D. Hoover. (1991). "Quality Control in the Development and Use of Performance Assessments." ***Applied Measurement in Education***, 4(4): 289-303.

Eagle, K.W., S. Allen, J. Hildreth, & T. Spiggle (1997). ***Goals 2000: Supporting State and Local Educational Improvement***. Washington, DC: Policy Studies Associates, Inc.

Education Commission of the States. (1999). ***Governing America's Schools: Changing the Rules***. Denver, CO: Author.

____. (1998). ***Designing and Implementing Standards-Based Accountability Systems***. Denver, CO: Author.

____. (1997). ***Education Accountability Systems in 50 States***. Denver, CO: Author.

Education Week. (2000)**. *Quality Counts 2000 Who Should Teach?, Vol. XIX, No. 18***. Bethesda, MD: Author.

____. (1999). ***Quality Counts '99 Rewarding Results, Punishing Failure, Vol. XVIII, No. 17***. Bethesda, MD: Author.

Elmore, R.F. (1998)

Elmore, R.F. (1996). "Getting to scale with successful educational practices." In S. H. Fuhrman & J. O'Day (Eds.), ***Reward and Reform: Creating educational incentives that work***. San Francisco: Jossey-Bass, pp. 294-329.

Elmore, R.F. (1996). "Getting to scale with successful educational practice." ***Harvard Educational Review***. 66(1): 1-26.

____. (1994). "Incentives for going to scale with effective practices: Some implications for federal policy and evaluation design." In U.S. Department of Education, ***Issues and Strategies in Evaluating Systemic Reform.*** Washington, DC: Author.

____. (1993). "The role of local school districts in instructional improvement." ***Designing Coherent Education Policy: Improving the System***. San Francisco, CA: Jossey-Bass Publishers, pp. 96-124.

Elmore, R.F., Abelmann, C.H., & S.H. Fuhrman. (1996). "The new accountability in state education reform: From process to performance" in ***Holding Schools Accountable: Performance-Based Reform in Education***. Washington, DC: The Brookings Institution.

Elmore, R.F., & M.W. McLaughlin (1988). ***Steady work: Policy, practice and the reform of American education***. Santa Monica, CA: RAND Corporation.

Elmore, R.F. with D. Burney. (1997). ***Investing in Teacher Learning: Staff Development and Instructional Improvement in Community School District #2, New York City***. New York, NY: National Commission on Teaching & America's Future and the Consortium for Policy Research in Education.

Epstein, J.L. (1991). "Effects on student achievement of teachers' practices of parent involvement." ***Advances in Reading/Language Research***. 5: 261-276.

Epstein, J.L., Clark, L.A., & F. Van Voorhis. (1999). "Two-year patterns of state and district leadership in developing programs of school, family, and community partnerships." Presented at the annual meeting of the American Educational Research Association.

Epstein, J.L. & Sanders, M.G. (1996)."School, family, community partnerships: Overview and new directions" In ***Education and Sociology: An Encyclopedia***, eds., D.L. Levinson, A.R. Sadovnik, P.W. Cookson, Jr., New York: Garland Publishing.

Evertson, C.M. (1986). "Do teachers make a difference? Issues for the eighties." ***Education and Urban Society***, 18(2).

Fine, C. (1994). ***Professional development: Changing times***. NCREL Policy Briefs: Report 4. Oak Brook, IL: The North Central Regional Education Laboratory (NCREL).

Finn, C.E. and M.J. Petrilli. (2000). ***The state of state standards 2000***. Washington, DC: Thomas B. Fordham Foundation.

Finn, C.E., M. Kanstoroom, and M.J. Petrilli. (1999). ***The Quest for Better Teachers: Grading the States***. Washington, DC: Thomas B. Fordham Foundation.

Finn, C.E., M. Petrilli, & G. Vanourek. (1998). ***The state of state standards***. Washington DC: Thomas B. Fordham Foundation.

Finn, J.D., & C.M. Achilles, C.M. (1990). "Answers and questions about class size: A statewide experiment." ***American Educational Research Journal***, *27*(3): 557-577.

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

Floden, R.E., M. Goertz, & J. O'Day. (1995). "Capacity Building in Systemic Reform." ***Phi Delta Kappan***. pp. 19-21.

Forrester, J. (1968). ***Principles of Systems.*** Pegasus.

Fuhrman, S.H. (1999). ***The New Accountability***. (CPRE Policy Brief No. RB-27). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

____. (1994). "New accountability systems and evaluating systemic reform. In Issues and strategies in evaluating systemic reform." Paper prepared for the U.S. Department of Education Office of the Under Secretary Planning and Evaluation Service.

____. (1994). ***Politics and Systemic Education Reform***. CPRE Policy Brief, No. RB-12-4. Philadelphia: Consortium for Policy Research in Education.

____. (1994). ***Challenges in Systemic Education Reform***. CPRE Policy Brief, No. RB-14-9. Philadelphia: Consortium for Policy Research in Education.

____. (1994). "Legislatures and education policy." In R. F. Elmore & S. H. Fuhrman (Eds.) ***The Governance of Curriculum***. 1994 Yearbook of the Association for Supervision and Curriculum Development. Alexandria, VA: ASCD.

____. (1993). "The politics of coherence." In S. H. Fuhrman (Ed.), ***Designing Coherent Education Policy: Improving the System***. San Francisco: Jossey-Bass.

Fuhrman, S.H. and D. Massell (1992). ***Issues and Strategies in Systemic Reform***. New Brunswick, NJ: Consortium for Policy Research in Education.

Fuhrman, S.H., & R.F. Elmore (1992). ***Takeover and Deregulation: Working Models of New State and Local Regulatory Relationships***. New Brunswick, NJ: Consortium for Policy Research.

Fuhrman, S.H., & R.F. Elmore (1990). "Understanding local control in the wake of state education reform." ***Educational Evaluation and Policy Analysis***, 12(1): 82-96.

Fullan, M.G. (1994a). "Coordinating Top-Down and Bottom-Up Strategies for Educational Reform." In S. Fuhrman & D. Elmore (Eds.), ***Governing Curriculum***. Alexandria, VA: ASCD, pp. 186-202.

____. (1994b). "Turning systemic thinking on its head." Paper prepared for the US Department of Ed, Office of the Under Secretary, Planning and Evaluation Service.

____. (1991). ***The New Meaning of Educational Change***. New York, NY: Teacher's College Press.

Fullan, M.G. & Miles, M.B. (1992)."Getting Reform Right: What Works and What Doesn't." ***Phi Delta Kappan***, 73(10): 744-52.

Fullan, M., & Stiegelbauer, S. (1991). ***The new meaning of educational change***. New York: Teachers College Press.

Gardner, H. (1992). "Assessment in context: The alternatives to standardized testing." In B. Gifford and M.C. O'Connor (Eds.) ***Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction.*** Boston, MA: Kluwer.

Garet, M.S., B.F. Birman, A.C. Porter, L. Desimone, & R. Herman, with K.S. Yoon. (1999). ***Designing Effective Professional Development: Lessons from the Eisenhower Program***. Washington, DC: American Institutes for Research.

Garlington, J. (1991). ***Helping Dreams Survive: The Story of a Project Involving African American Families in the Education of their Children***. Washington, D.C.: National Committee for Citizens in Education (ED 340 805).

Gearhart, M., J.L. Herman, E.L. Baker, and A.K. Whittaker. (1993). ***Whose Work Is It? A Question for The Validity of Large-scale Portfolio Assessment.*** Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Glaser, R., & E. Silver (1994). ***Assessment, Testing, and Instruction: Retrospect and Prospect.*** Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Glass, G. (1987). ***Standards of Competence: A Multi-site Case Study of School Reform.*** Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Goertz, M., D. Massell, & T. Chun. (1998). ***District Response to State Accountability Systems***. Paper presented at the Annual Meeting of the Association for Public Policy Analysis and Management. New York, NY.

Goertz, M., R. Floden & J. O'Day. (1996). ***The Bumpy Road to Education Reform***. CPRE Policy Brief RB-20-June 1996. Philadelphia: Consortium for Policy Research in Education.

____. (1996). ***Systemic Reform: Studies in Education Reform***. Washington, D.C.: Office of Educational Research and Improvement, U.S. Department of Education.

Grant, S.G., P.L. Peterson, & A. Shojgreen-Downer. (1996). "Learning to teach mathematics in the context of systemic reform." ***American Educational Research Journal***, 33: 502-541.

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

Haladyna, T.M., S.B. Nolen, & N.S. Haas. (1991). "Raising standardized achievement test scores and the origins of test score pollution." ***Educational Researcher***, 20(5), 2-7.

Hannaway, J. & K. Kimball. (Forthcoming). ***Big Isn't Always Bad: School District Size, Poverty, and Standards-Based Reform***. Washington, DC: The Urban Institute.

Hannaway, J. with K. Kimball. (1997). ***Reports on Reform from the Field: District and State Survey Results. Final Report*****.** Washington DC: The Urban Institute.

Hannaway, J., et al. (1996). ***Study of Federal Efforts to Assist States in Education Reform.*** Document review submitted to the U.S. Department of Education, Planning and Evaluation Service. Washington, DC: The Urban Institute.

Hanushek, E.A. & R.H. Meyer. (1996). "Comments on Chapters Two, Three, and Four" in ***Holding Schools Accountable: Performance-Based Reform in Education***. Washington, DC: The Brookings Institution.

Hatry, H.P. (1994). ***Development of Performance Indicator Systems for Evaluation of Systemic Reform***. Washington, DC: The Urban Institute.

Hatry, H.P., & M. Kopczynski. (1996). ***A Process for Regular Assessment of Progress in Educational Reform by U.S. Department of Education Regional Service Teams, Vol. I & II***. Washington DC: The Urban Institute.

Henderson, A. (Ed.). (1987). ***The Evidence Continues to Grow: Parent Involvement Improves Student Achievement.*** Columbia, MD.: National Committee for Citizens in Education.

Herman, J.L. (1997). "Assessing New Assessments: How Do They Measure Up?" ***Theory into Practice***, 36(4).

____. (1997). ***Large-scale assessment in support of school reform: Lessons in the search for alternative measures.*** (CSE Technical Report 446). Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

____. (1992). ***Accountability and alternative assessment: Research and development issues.*** Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Herman, J.L., & S. Golan (1993). "Effects of standardized testing on teaching and schools." ***Educational Measurement: Issues and Practice,*** 12(4): 20-25, 41-42.

____. (1991). ***Effects of standardized testing on teachers and learning: Another look.*** Los Angeles, CA: University of California, Center for the Study of Evaluation.

Janas, M. & Gurganus, S. "Meeting the diverse needs of special educators through a collaborative model for science reform." ***Journal of Staff Development***, 14(4), 1993.

Johnston, R.C. & J.L. Sandham. (1999). "States Increasingly Flexing Their Policy Muscle." ***Education Week***, April 14, 1999.

Kahle, J.B. (1999). "Discovering from discovery: The evaluation of Ohio's Systemic Initiative." Paper presented at the Fourth Annual NISE Forum, Arlington, VA.

____. (1998). ***Reaching Equity in Systemic Reform: How Do We Assess Progress and Problems?*** Research Monograph No. 9. Madison, WI: NISE.

Kellaghan, T., & G. Madaus (1991). "National testing: Lessons for America from Europe." ***Educational Leadership,*** 49(3): 97-93.

Kennedy, M. (1998). "Form and substance in inservice teacher education" (Research Monograph #13). University of Wisconsin, Madison: National Institute for Science Education.

Kentucky Department of Education (KDE). (1994). ***Kentucky Instructional Results Information System, 1992-93 Technical Report***. Frankfort, KY: Author.

Knapp, M.S. (1997). "Between systemic reforms and the mathematics and science classroom: The dynamics of innovation, implementation, and professional learning." ***Review of Educational Research***, 67(2), 227-266.

Knapp, M.S., & B.J. Turnbull (1990). ***Better schooling for the children of poverty: Alternatives to conventional wisdom. Study of academic instruction for disadvantaged children: Volume I, Summary.*** Washington, DC: U.S. Department of Education.

Koretz, D., Deibert, E., & Stecher, B. (1994). ***The Vermont portfolio assessment program: Findings and implications***. Washington, DC: RAND Institute on Education and Training.

Koretz, D., B. Stecher, & E. Deibert. (1993). ***The Reliability of Scores From the 1992 Vermont Portfolio Assessment Program.*** Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Ladd, H.F. (1996). "Introduction" in ***Holding Schools Accountable: Performance-Based Reform in Education***. Washington, DC: The Brookings Institution.

Ladd, H.F. ed. (1996). ***Holding schools accountable: Performance-based reform in education***. Washington, DC: The Brookings Institution.

Laguarda, K.G. (1998). ***Assessing the SSIs' impacts on student achievement: An imperfect science***. Menlo Park, CA: SRI International and Washington, DC: Policy Studies Associates.

LaLonde, L. (1986). "Evaluating the econometric evaluations of training programs with experimental data." ***American Economic Review,*** 76: 604-620.

Lee, J. (1996). "Multilevel linkages of state education reform to instructional practices." Presented at the April 1996 annual meeting of the American Educational Research Association.

Leighton, M.S., Mullens, J.E., Turnbull, B., Weiner, L.K. & Williams, A.S. (1995). ***Measuring instruction, curriculum content, and instructional resources: The status of recent work. Working paper series***. Washington, DC: National Center for Education Statistics.

Leighton, et al. (2000). ***Pulling in the Same Direction: Goals 2000 and State Standards-Based Reform***. Draft Executive Summary. Washington, DC: Policy Studies Associates, Inc.

LeMahieu, P., D. Gitomer, & J. Eresh (1994). ***Portfolios beyond the classroom: Data quality and qualities.*** Princeton, NJ: Educational Testing Service.

LeTendre, M.J. (1991). "The continuing evolution of a federal role in compensatory education." ***Educational Evaluation and Policy Analysis***, 13(4): 328-334.

Lignon, G. (1991). "Models for identifying Chapter 1 schools for improvement." ***Educational Evaluation and Policy Analysis***, 13(4), 389-393.

Lindle, J.C. (1994). "Kentucky's reform opens doors to family involvement." ***Dimensions of Early Childhood***, 22(2): 20-22.

Linn, R.L. (1998). ***Assessments and accountability***. (CSE Technical Report 490). Los Angeles, CA: University of California, Los Angeles, Center for the Study of Evaluation.

____. (1994). "Performance assessment: Policy promises and technical measurement standards." ***Educational Researcher***, 23(9): 4-14.

Linn, R.L. (1993). "Educational assessment: Expanded expectations and challenges." ***Educational Evaluation and Policy Analysis***, 15(1): 1-16.

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

Linn, R.L., E.I. Baker, & S.B. Dunbar (1991). "Complex, performance-based assessment: Expectations and validation criteria." ***Educational Researcher***, 20(8): 15-21.

Linn, R.L. & J.L. Herman. (1997). ***Standards-led assessment: Technical and policy issues in measuring school and student progress***. (CSE Technical Report 426). Los Angeles, CA: University of California, Los Angeles, Center for the Study of Evaluation.

Linn, R.L. & E.L. Baker. (1996). "Can performance-based assessments be psychometrically sound?" in J.B. Baron and D.P. Wolf (Eds.), ***Performance-based Student Assessment: Challenges and Possibilities, 87th Yearbook of the National Society for the Study of Education***. Chicago, IL: University of Chicago Press, Part 1, pp. 84-103.

Little, J.W. (1993). "Teachers' professional development in a climate of educational reform." ***Educational Evaluation and Policy Analysis***. 15(2): 129-151.

Little, J.W. & M.W. McLaughlin, eds., (1993). ***Teachers' work: Individuals, colleagues, and contexts***. New York: Teachers College Press.

Loucks-Horsley, S. (1998). "The Role of Teaching and Learning in Systemic Reform: A Focus on Professional Development." ***Science Educator***, 7(1): 1-6.

Loucks-Horsley, S., Hewson, P.W., Love, N., & K.E. Stiles. (1998). ***Designing Professional Development for Teachers of Science and Mathematics***. Thousand Oaks, CA: Corwin Press.

Madaus, G.F. (1985). "Public policy and the testing profession: You've never had it so good." ***Educational Measurement: Issues and Practice***, 4(4): 5-11.

Madaus, G.F., et al. (1992). ***The Influence of Testing on Teaching Math and Science in Grades 4-12***. Chestnut Hill, MA: Boston College.

Madden, N.A., R.E. Slavin, N.L. Karweit, L. Dolan, & B.A.Wasik (1990). ASuccess for All: Effects of variations in duration and resources of a schoolwide elementary restructuring program." Paper presented at the annual convention of the American Educational Research Association, Boston, MA.

Marzano, R.J. & J.S. Kendall with B.B. Gaddy. (1999). ***Essential Knowledge: The Debate Over What American Students Should Know.*** Aurora, CO: Mid-continental Educational Research Laboratory.

Marzano, R.J. & J.S. Kendall. (1998). ***The Status of State Standards.*** Aurora, CO: Mid-continent Research for Education and Learning, Inc.

Massell, D. (1998). ***State Strategies for Building Local Capacity: Addressing the Needs to Standards-Based Reform***. CPRE Policy Brief. RB-25-July 1998.

Massell, D., M. Kirst, & M. Hoppe (1997). ***Persistence and Change: Standards-based Reform in Nine States***. (CPRE Research Series Report #37). Philadelphia, PA: Consortium for Policy Research in Education.

Mayer, D.P. (1999). "Measuring instructional practice: Can policymakers trust survey data?" ***Educational Evaluation and Policy Analysis***, 21: 29-45.

McDonnell, L.M. (1994). ***Policymakers' Views of Student Assessment***. CSE Technical Report 378. Graduate School of Education & Information Studies, Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

McGregor, G., A. Halvorsen, D. Fisher, I. Pumpian, B. Bhaerman, & C. Salisbury. (1998). ***Professional Development for All Personnel in Inclusive Schools***. Consortium on Inclusive Schooling Practices Issue Brief 3(3).

McLaughlin, M.J., K. Henderson, & L.M. Rhim. (1998). ***Snapshots of Reform: How Five Local Districts Are Interpreting Standards-Based Reform for Students with Disabilities.*** Alexandria, VA: Center for Policy Research on the Impact of General and Special Education Reform.

McLaughlin, M.W. & Talbert, J.E. (1993). ***Contexts that matter for teaching and learning***. Stanford, CA: Stanford University.

McLaughlin, M.W., L.A. Shepard, & J.A. O'Day (1995). ***Improving education through standards-based reform. A Report by the National Academy of Education Panel of Standards-Based Education Reform***. Stanford, CA: Stanford University, The National Academy of Education.

Mitchell, A. (In publication). "Historical Trends in Federal Education Policies that Target Students Placed At-Risk." In Mavis Sanders, (Ed.), ***The Education of Students Placed at Risk***.

Mitchell, A., & J. Raphael (1999). ***Goals 2000: Case Studies of Exemplary Districts.*** Washington, DC: The Urban Institute.

Moles, O.C. (1993). "Collaboration between Schools and Disadvantaged Parents: Obstacles and Openings." in Chavkin, N.F. (Ed.). ***Families and Schools in a Pluralistic Society.*** New York: State University of New York Press.

National Academy of Education's Panel on Standards-based Reform (1995). ***Improving Education Through Standards-based Reform***. Stanford, CA: Stanford University.

National Alliance of Business (1998). "Business, educators find power in Baldridge to improve schools." ***Work America***. 15(4).

National Assessment of Educational Progress (1998). ***Report in Brief: NAEP 1996 Trends in Academic Progress***. Washington, DC: National Center for Educational Statistics.

____. (1997). ***NAEP 1996: Trends in Academic Progress,*** Washington DC: National Center for Educational Statistics.

____. (1995). ***NAEP 1994 Reading: A First Look -- Findings from the National Assessment of Educational Progress.*** Washington DC: National Center for Educational Statistics.

National Center for Education Statistics. (1999). ***Status of Education Reform in Public Elementary and Secondary Schools: Teachers' Perspectives***. Washington, DC: US Department of Education.

National Center on Educational Outcomes. (1997). ***1997 State Special Education Outcomes: A Report on State Activities During Educational Reform***. Minneapolis: Author.

National Commission on Excellence in Education (1983). ***A Nation at Risk: The Imperative for Educational Reform.*** Washington, DC: US Government Printing Office.

National Council of Teachers of Mathematics (1988). ***Curriculum and Evaluation Standards for School Mathematics***. Reston, VA: Author.

National Research Council. (1999). ***Testing, Teaching, and Learning: A Guide for States and School Districts.*** Elmore, R.F. and Rothman, R. (eds.) Washington, DC: National Academy Press.

National Science Foundation. (1998). ***Infusing Equity in Systemic Reform: An Implementation Scheme***. Arlington, VA: Author.

Nettles, S.M. (1991). "Community Contributions to School Outcomes of African-American Students." ***Education and Urban Society***, 24(1): 132-147.

____. (1991). "Community Involvement and Disadvantaged Students: A Review." ***Review of Educational Research***, 61(3): 379-406.

O'Day, J.A. & Smith, M.S. (1993). "Systemic Reform Educational Opportunity." in S. H. Fuhrman (Ed.), ***Designing Coherent Educational Policy: Improving the System***. San Francisco, CA: Jossey-Bass, 1993, pp. 250-312.

Oakes, J. (1990). ***Multiplying Inequalities: The Effects of Race, Social Class, and Tracking on Opportunities to Learn Mathematics and Science.*** Santa Monica, CA: RAND.

____. (1986). ***Educational Indicators: A Guide for Policymakers***. (CPRE Occasional Paper No. OPE-01). New Brunswick, NJ: Rutgers University, Center for Policy Research in Education.

____. (1985). ***Keeping Track: How Schools Structure Inequality.*** New Haven, CT: Yale University Press.

Odden, A.R. (1991). "The evolution of education policy implementation." In A.R. Odden (Ed.) ***Educational policy implementation***. Albany: State University of New York Press.

Office of Technology Assessment (1992). ***Testing in American Schools: Asking the Right Questions***, Washington, DC: Author.

Olebe, M.G. & Others. (1992). "Consider the Customer." ***American School Board Journal***, 179(12): 52-55.

Oregon Department of Education. (1998). "Proposed continuous improvement model for assessing district effectiveness." Draft document.

Orlich et. al. (1993). "Seeking the link between student achievement and staff development." ***Journal of Staff Development***, 14(3).

Orr, L.L. (1998). ***Social Experiments: Evaluating Public Programs with Experimental Methods.*** Thousand Oaks, CA: Sage Publications Inc.

Page, R. (1995). "Who systematizes the systematizers? Policy and practice interactions in a case of state-level systemic reform." ***Theory Into Practice***, 34: 21-29.

Pellegrino, J.W. (1992). "Commentary: Understanding what we measure and measuring what we understand." In B. Gifford & M.C. O'Connor (Eds.) ***Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction.*** Boston, MA: Kluwer.

Plank, D.N., & W.L. Boyd. (1994). "Antipolitics, education, and institutional choice: The flight from democracy." ***American Educational Research Journal***, 31(2): 263-281.

Porter, A.C. (1994). "National Standards and School Improvement in the 1990s: Issues and Promise." *American Journal of Education*, 102(4), 421-449.

____. (1993). "Defining and measuring opportunity to learn." A paper prepared for the National Governors' Association.

Porter, A., R. Floden, D. Freeman, W. Schmidt, & J. Schwille. (1988). "Content Determinants in Elementary School Mathematics." In ***Perspectives on Research on Effective Mathematics Teaching. Vol. 1***. D.A. Grouws and T.J. Cooney, eds. Lawrence Erlbaum Associates, National Council of Teachers of Mathematics: Reston, VA.

Porter, A.C., M.W. Kirst, E.J. Osthoff, J.L. Smithson, & S.A. Schneider. ***Reform Up Close: A Classroom Analysis***. Madison, WI: Wisconsin Center for Education Research.

Puma, M.J., et al. (1997). ***Prospects: Final Report on Student Outcomes.*** Washington, DC: US Department of Education.

Resnick, L.B. & D.P. Resnick. (1992). "Assessing the Thinking Curriculum: New Tools for Educational Reform" in B.R. Gifford and M.C. O'Connor (Eds.), ***Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction***. Boston, MA: Kluwer Academic Publishers, pp. 37-75.

Rhine, S. (1995). "The challenge of effectively preparing teachers of limited-English proficient students." ***Journal of Teacher Education***, 46(5).

Ritter, P.L., Mont-Reynaud, R., & Dornbush, S.M. (1993). "Minority Parents and Their Youth: Concern, Encouragement, and Support for School Achievement" in Chavkin, N.F. (Ed.). ***Families and Schools in a Pluralistic Society***. New York: State University of New York Press.

Romberg, T.A., & T.P. Carpenter (1986). "Research on teaching and learning mathematics: Two disciplines of scientific inquiry." In M.C. Wittrock (Ed.). ***Handbook of Research on Teaching.***

Sanders, M.G. (1999). "Collaborating for Student Success: A Study of the Role of "Community" in Comprehensive School, Family, and Community Partnership Programs." Presented at the 1999 annual meeting of the American Educational Research Association.

Schlechty, P.C. (1993). "On the frontier of school reform with trailblazers, pioneers and settlers." ***Journal of staff development***, 14(4).

Seidman, I.E. (1991). ***Interviewing as Qualitative Research***. New York: Teachers College Press.

Shavelson, R.J., G.P. Baxter, & X. Gao. (1993). "Sampling variability of performance assessments." ***Journal of Educational Measurement,*** 30: 215-232.

Shavelson, R.J., G.P. Baxter, & J. Pine. (1992). "Performance assessments: Political rhetoric and measurement reality." ***Educational Researcher,*** 21(4): 22-27.

_____. (1991). "Performance assessment in science." ***Applied Measurement in Education,*** 4: 347-362.

Shepard, L. (1991). "Will national tests improve student learning?" ***Phi Delta Kappan***, 73(3): 232-238.

____. (1990). ***Inflated Test Score Gains: Is it Old Norms or Teaching to the Test?*** Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Shepard, L.A., et al. (1995). ***Effects of Introducing Classroom Performance Assessments on Student Learning.*** Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Shepard, L. & M.L. Smith (1988). "Escalating academic demand in kindergarten: Counterproductive policies." ***The Elementary School Journal,*** 89: 135-145.

Shields, P.M, J. Marsh, & N.E. Adelman (1998). ***The SSIs' Impacts on Classroom Practice***. Menlo Park, CA: SRI International.

Shields, P.M., T.B. Corcoran, & A. Zucker (1994). ***Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program***. Washington, DC: National Science Foundation.

Simon, H. (1969). ***The Sciences of the Artificial***. Cambridge, MA: MIT Press.

Slavin, R.E., & N.A. Madden, (1991). "Modifying the Chapter 1 program improvement guidelines to reward appropriate practices." ***Educational Evaluation and Policy Analysis***, 13(4): 369-379.

Sleeter, C. (1992). "Restructuring schools for multicultural education." ***Journal of Teacher Education***, 43(2).

Smith, M.L. (1991). "Meanings of test preparation." ***American Education Research Journal***, 28(3), 521-542.

Smith, M.L., & C. Rottenberg (1991). "Unintended consequences of external testing in elementary schools." ***Educational Measurement: Issues and Practice,*** 10(4): 7-11.

Smith, M.S., & J. O'Day (1991). "Systemic school reform." In S.H. Fuhrman & B. Malen (Eds.), ***The Politics of Curriculum and Testing, 1990 Yearbook of the Politics of Education Association***. London and Washington, DC: Falmer Press, 233-267.

Smylie, M.A. (1996). "From Bureaucratic Control to Building Human Capital: The Importance of Teacher Learning in Education Reform." ***Educational Researcher***, 25(9): 9-11.

Solomon, R.P. (1995). "Beyond prescriptive pedagogy: Teacher inservice education for cultural diversity." ***Journal of Teacher Education***, 46(4).

Spillane, J.P. (n.d.). "Cognition and Policy Implementation: District Policy-makers and the Reform of Mathematics Education." ***Cognition and Instruction***. (under review).

____. (1999). "External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment." ***Journal of Curriculum Studies***. 31(2); 143-175.

____. (1999). "State and Local Government Relations in the Era of Standards Based Reform: Standards, State Policy Instruments, and Local Instructional Policy-making." ***Educational Policy***. 13(4).

____. (1998). "A Cognitive Perspective on the Role of the Local Educational Agency in Implementing Instructional Policy: Accounting for Local Variability." ***Education Administration Quarterly***. 34(1): 31-57.

____. (1998). "State Policy and the Non-Monolithic Nature of the Local School District: Organizational and Professional Considerations." ***American Educational Research Journal***. 35(1): 33-63.

____. (1996). "School Districts Matter: Local Educational Authorities and State Instructional Policy." ***Educational Policy***. 10(1): 63-87.

____. (1994). "How districts mediate between state policy and teachers' practice." In R. E. Elmore & S. H. Fuhrman (Eds.), ***The Governance of Curriculum. 1994 Yearbook of the Association for Supervision and Curriculum Development***. Alexandria, VA: ASCD, 167-185.

Spillane, J.P. & C.L. Thompson. (1997). "Reconstructing Conceptions of Local Capacity: The Local Education Agency's Capacity for Ambitious Instructional Reform." ***Educational Evaluation and Policy Analysis***, 19(2): 185-203.

Spillane, J.P. & N.E. Jennings. (1997). "Aligned Instructional Policy and Ambitious Instructional Reform from the Classroom Perspective." ***Teachers College Record***. 98: 3.

Spillane, J.P. & J.S. Zueli. (1999). "Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms." ***Educational Evaluation and Policy Analysis***, 21: 1-28.

Stiegelbauer, S.M. (1994). "Change has changed: Implications for implementation of assessments from the organizational change literature." In ***Systemic Reform: Perspectives on Personalizing Education***. http://www.ed.gov/pubs/EdReformStudies/SysReforms/stiegel1.html, January 20, 1999.

Stiggins, R.J. (1991). "Facing the challenge of a new era of educational assessment." ***Applied Measurement in Education***, 10(2): 35-39.

Stringfield, S., S.H. Billig, & A. Davis (1991). "Chapter 1 program improvement: Cause for cautious optimism and a call for much more research." ***Educational Evaluation and Policy Analysis***, 13(4): 399-406.

Swanson, J. (1995). "Systemic reform in the professionalism of educators***." Phi Delta Kappan***.

Tangri, S.S. & Moles, O. (1987). "Parents and the Community" in V. Richardson-Koehler (Ed.), ***Educator's Handbook***. New York: Longman.

Task Group on Systemic Reform. (1993). ***Building Capacity for Systemic Reform***. Washington, DC: US Department of Education, OERI.

Taylor, C. (1994). "Assessment for measurement or standards: The peril and promise of large-scale assessment reform." ***American Educational Research Journal***, 31(2): 231-262.

Thompson, S., R. Erickson, M. Thurlow, J. Ysseldyke, & S. Callender. (1999). ***Status of the States in the Development of Alternate Assessments***. National Center on Educational Outcomes.

Turnbull, B., M. Welsh, C. Heid, W. Davis, & A. Ratnofsky (1999). ***The Longitudinal Evaluation of School Change and Performance (LESCP) in Title I Schools***. Rockville, MD: Westat, Inc.

U.S. Department of Education. (1998). ***Goals 2000: Reforming Education to Improve Student Achievement***. Goals 2000 Report to Congress, Washington, DC: Author.

____. (1996a). ***Goals 2000: Educate America Act, October 1996 Update.*** Washington, DC: Author.

____. (1996b). ***Mapping Out the Assessment of Title 1: The Interim Report***. Washington, DC: Author.

____. (1995). ***The Dwight D. Eisenhower Professional Development Program***. Office of Elementary and Secondary Education. Washington, DC: Author.

____. (1993a). ***Building Capacity for Systemic Reform***. Washington, DC: Author.

____. (1993b). ***Reinforcing the Promise, Reforming the Paradigm: Report of the Advisory Committee on Testing in Chapter 1.*** Washington, DC: Author.

____. (1993c). ***Reinventing Chapter 1: the Current Chapter 1 Program and New Directions Final Report of the National Assessment of the Chapter 1 Program***. Washington, DC: Author.

____. (1992). ***National Assessment of the Chapter 1 Program: The Interim Report***. Washington, DC: Author.

U.S. General Accounting Office. (1998a). ***Goals 2000: Flexible Funding Supports State and Local Education Reform***. Washington, DC: Author.

____. (1998b). ***Grant Programs: Design features shape flexibility, accountability, and performance information***. Washington, DC: Author.

____. (1993a). ***Systemwide Education Reform: Federal Leadership Could Facilitate District-level efforts.*** Washington, DC: Author

____. (1993b). ***Chapter 1 Accountability: Greater Focus on Program Goals Needed***. Washington, DC: Author.

Vinovskis, M.A. (1996). "An analysis of the concept and uses of systemic educational reform." ***American Educational Research Journal***, 33(1), 53-85.

Von Bertalanffy, L. (1976). ***General Systems Theory***. George Braziller.

Walberg, H.J. (1984). "Improving the Productivity of America's Schools." ***Educational Leadership***, 41(8): 19-27.

Wang, M.C., G.D. Haertel, & H.J. Walberg (1993). "Toward a knowledge base for school learning." ***Review of Educational Research***, 63(3): 249-294.

Weatherly, R., & M. Lipsky (1978). "Street-level bureaucrats and institutional innovation: Implementing special education reform." ***Harvard Education Review,*** 47: 171-197.

Webb, N. (1999). ***Alignment of Science and Mathematics Standards and Assessments in Four States***. Madison, WI: National Institute for Science Education; Washington, DC: Council of Chief State School Officers.

____. (1997). ***Determining Alignment of Expectations and Assessments in Mathematics and Science Education***. National Institute for Science Education (NISE) Brief, 1(2). Madison, WI: National Institute for Science Education.

____. (1999). ***Evaluation of Systemic Reform: Confronting the Challenges***. Paper presented at the 1999 Fourth Annual NISE Forum.

Weiner, N. (1986). ***Norbert Weiner: Cybernetics, Science, and Society (Collected Works)***. Cambridge, MA: MIT Press.

Weiss, I. (1999). "Evaluating systemic reform: A complex endeavor." Paper presented at the 1999 Fourth Annual NISE Forum.

Weller, L.D., & Wellner, S.J. (1997). "Using Deming's Continuous Improvement Model to Improve Reading." ***NASSP Bulletin***, 81(589): 78-85.

Wiggins, G. (1989). "Teaching to the (authentic) test." ***Educational Leadership***, 41: 47.

Winfield, L.F., & M.D. Woodard (1994). ***Assessment, Equity, and Diversity in Reforming America's Schools.*** Los Angeles, CA: National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Wood, F.H. & Thompson, S.R. (1993)."Assumptions about staff development based on research and best practice." ***Journal of Staff Development.*** 4.

Ysseldyke, J.; Krentz, J.; Elliott, J.; Thurlow, M.; Thompson, S.; & Moore, M. (1998). ***NCEO Framework for Educational Accountability: Post-School Outcomes***. Minneapolis, MN: National Center on Education Outcomes.

Ysseldyke, J.E., M.L. Thurlow, K.L. Langenfeld, J.R. Nelson, E. Teelucksingh, & A. Seyfarth. (1998). ***Educational Results for Students with disabilities: What Do the Data Tell Us? Technical Report No. 23.*** Minneapolis: National Center on Educational Outcomes.

Zucker, A.A. & P.M. Shields (1997). ***SSI Strategies for Reform: Preliminary Findings from the Evaluation of NSF's SSI Program.*** Menlo Park, CA; SRI International.

Zucker, A.A., P.M. Shields, N.E. Adelman, T.B. Corcoran, M.E. Goertz. (1998). ***Statewide Systemic Initiatives Program***. Menlo Park, CA: SRI International; Arlington, VA: National Science Foundation.

Chapter VIII: Appendices

Appendix A: Criteria suggested in other publications for describing systemic reform

Appendix B

The shift in emphasis to "high stakes" testing of large numbers of students, especially using standardized norm-referenced tests, has become one of the most hotly debated issues in American education today.

Criticisms of Norm-Referenced Testing

Critics have identified a variety of concerns about the detrimental effects of large-scale student testing on both schools and students. Some of the important themes raised by these critics include the following:

- ***Standardized tests do not reflect current thinking about human development.*** Opponents of multiple-choice tests have argued that these assessments have forced instruction to become less contextualized in order to match the focus of most standardized tests on discrete facts. This narrowing of instruction runs counter to how theorists now believe students actually gain knowledge, which is not through the structured and progressive accumulation of discrete skills and/or facts, but, rather, through a process where humans "construct" knowledge within meaningful contexts (Resnick & Resnick, 1992; Gardner, 1992; Herman, 1997).

- ***Standardized tests have determined the curriculum content and the form of instruction.*** Several researchers have found that high-stakes standardized testing, particularly with the added link to teacher salaries and school prestige, has led many teachers to focus on the narrow skills included in the multiple-choice tests (Darling-Hammond & Wise, 1985; Haladyna, Nolen, & Haas, 1991; Koretz, et al., 1991; Shepard, 1991; Madaus, et al., 1992). Teachers have also been found to spend significant amounts of time preparing students for the tests -- in some instances, weeks of class time was taken up with drill and practice preparation (Corbett & Wilson, 1991; Dorr-Bremme & Herman, 1986; Herman & Golan, 1991; Kellaghan & Madaus, 1991; Shepard, 1991; Smith & Rottenberg, 1991). Moreover, teachers have been found to de-emphasize subjects that were ***not*** part of the testing agenda, especially science and social studies (Darling-Hammond & Wise, 1985; Shepard, 1991), and this negative impact on instructional content was found to be worse for students in schools serving higher proportions of disadvantaged children (Herman & Golan, 1993).

- ***There is a strong incentive to "teach to the test."*** Cannell (1987), in a famous study (later substantiated by Linn, Graue, & Sanders, 1990), found that almost all states reported that their students were scoring above the national norm sample. Subsequent research by Shepard (1990) indicated that the reason for this finding was that teachers were directly teaching to the test. Even worse, the practice had become so

institutionalized that high test scores were found not to translate into higher performance as measured by other types of assessments.

- ***Standardized testing has sustained "student tracking."*** The use of tests to identify "strong" and "weak" students has become a hotly debated topic. Opponents argue that it disproportionately shunts poor and minority students into instructional "tracks" where they receive poor quality instruction focused on the remediation of basic skills (Oakes, 1990), and where they receive less actual instruction (Oakes, 1985). Both effects have prevented them from "closing the gap" between themselves and their more advantaged peers (Puma, et al., 1997).

- ***Standardized tests can be misused by teachers and schools.*** Test results can, for example, be used to make inappropriate retention decisions (especially for low-achieving students) because retained students will naturally score higher on standardized tests when they are re-tested the following year (Shepard & Smith, 1988; Slavin & Madden, 1991). Slavin & Madden (1991), have further suggested that: (1) there can be a disincentive to invest in preschool, kindergarten, and 1st-grade programs because gains achieved at this level would be ignored and it would make it harder to demonstrate later increases in student achievement; (2) there is a risk that schools could inflate their average scores by excluding students "at the lower end" and, (3) because of the problem of "regression to the mean" schools can mis-interpret gains (i.e., retained students will generally score higher when re-tested).

- ***Standardized tests can negatively affect teachers***. The incentive to create test-directed instruction tends to degrade and de-skill teachers (Rottenberg & Smith, 1990). When instruction is tightly dictated from the outside, teachers have far less incentive to hone their skills, and are unlikely to seek new and innovative ways to teach their students.

- ***Most standardized testing does not provide useful information to teachers.*** At the classroom level, teachers do not find the data from large-scale standardized tests informative for diagnosing how students learn (Pellegrino, 1992) for at least two reasons: 1) the tests are usually given at the end of the year and reported after the school year has ended; and, 2) most standardized tests do not decompose the student's strengths and weaknesses to a level of detail that can allow teachers to decide on the best instructional strategy.

In response to many of these criticisms, a new line of tests have been developed in recent years, many of them closely tied to the systemic reform movement. But, "Norm-referenced assessment is not going away, probably because it provides national comparative data....Policymakers want to be ensured that their students are doing as well or better than students in other states ... [but] states are using a combination of assessment

types to meet the many types of content standards and assessment purposes" (CCSSO, 1998).

The Move to Alternative Tests

The new types of assessments seek to measure students on actual performances of what they can do, as opposed to trying to measure factual information that they know (Wiggins,1989). These new assessments go by many names including "alternative," "authentic," and "performance," and range across a variety of different activities including open-ended questions on a test, conducting a science experiment, writing an essay, working out the solution to a math problem, portfolio collections of students' work, and extended year-long projects. The key characteristic of this new approach is that students are required to "construct" their answers by bringing their knowledge and experience to bear to solve realistic problems (Herman, 1997).

These new approaches have caught on (CCSSO, 1998; Taylor, 1994; Baron 1991; Stiggins, 1991), largely on the basis of arguments made by proponents that, "Alternative forms of assessment...can adequately reflect today's educational goals and, if properly used, serve as positive tools in creating schools truly capable of teaching students to think" (Resnick & Resnick, 1992). In effect, proponents argue that "...things will be different. Better tests -- performance measures aimed at assessing higher-order learning goals -- will ensure student learning by redirecting instruction toward more challenging content (Shepard, 1991).

Despite the growing popularity of these new forms of assessment, many have worried about the validity and reliability of these new types of tests (Shavelson, Baxter, & Pine, 1992; Linn, Baker, & Dunbar, 1991). For example, Linn (1994), in a broad review of the recent literature and debate about student performance-based assessments, identifies a variety of concerns that need to be considered in using these types of tests. First, there are significant *political* issues related to their development and use as various groups of activists have mounted strong opposition to "outcomes-based" education, especially when parents and constituency groups do not clearly understand the intent of the testing,

as evidenced by the recent cancellation of the testing programs in California and Arizona. Second, there are a variety of *technical* issues that surround the development of new types of assessments:

- ***Validity.*** At a minimum, an assessment should reflect the skills and abilities that it is intended to measure. This is particularly difficult challenge as demonstrated by the problems that many states have encountered in trying to develop standards and curriculum frameworks that are expected to guide the development of the tests**.**

- ***Generalizeability.*** There is also evidence to suggest that scores on performance assessments may not correlate well with other types of presumably similar assessments in the same subject area (Shavelson, Baxter, & Pine, 1991; Gearhart, et al., 1993). Part of the difficulty arises from the number of tasks one needs to get a reliable estimate of an individual's performance in a particular area, i.e., the "sampling" of a small number of tasks (or portfolio items) and then using these to generalize to an assessment of overall student performance, particularly with the small number of classifications typically used (e.g., advanced proficient, not proficient). Many researchers have examined this issue and the results appear to indicate that an average of about 15-17 tasks is required (Herman, 1997). And, according to Shavelson, et al. (1993) the problem is further compounded when one wants to cover an entire subject area (mathematics) that can be comprised of multiple topics, each of which would require multiple tasks to obtain a valid assessment. The amount of time required to complete even a single subject performance assessment could, therefore, spread across several days of valuable school time.

  Related to this issue is the "breadth of interpretation," i.e., the extent to which test results are used to make broad judgements about an individual student, teacher, or school. For example, performance on a particular set of tasks could be used (often incorrectly) to make more general conclusions about a student's broader problem-solving ability, and likelihood of future life success

- ***Test reliability.*** The use of new performance-based assessments has also raised serious concerns about the reliability of the testing procedures given the high degree of difficulty associated with maintaining high rates of inter-rater reliability. A reliable test should give similar results at different times (e.g., if I weigh an object today and again next week, I should get the same results, barring any change in the object itself, of course).

  The problem with the new performance assessments is that they require human judgement for scoring raising concerns about the extent to which multiple raters can agree on the same score. The evidence indicates that this challenge will not be easy to overcome, but the problem may be tractable. First, with well defined rubrics, systematic procedures, and a high degree of rater training, these types of tests can yield reliable scoring of open-ended questions (Herman, 1997). Second, the results of

the Iowa Tests of Basic Skills writing assessment have shown the feasibility of achieving inter-rater reliabilities in the range of 0.9. Similarly, the Pittsburgh portfolio assessments yielded reliabilities that ranged from 0.84 to 0.87, but this experience also showed the importance of getting the raters ready to reach a consensus on scoring over a period of time (Herman, 1997). On the downside, the results of large-scale performance assessments in Vermont and Arizona were not publicly released due to concerns about the reliability of the test scores (Herman, 1997; Koretz, et al., 1993).

- ***Fairness***. In addition to worries about the reliability of the scoring procedures, there are concerns with the fairness of the evaluation process, especially given the subjective nature of the judgements being made by the raters. Results from the Pittsburgh portfolio assessments (LeMahieu, et al., 1994) indicate that although women raters scored higher than their male counterparts, men and women treated boys and girls the same. Similarly, there were no differential effects due to the race/ethnicity of the student and the rater.

  The use of portfolios can bring an additional set of concerns. First, there can be significant variation in the amount of time students (and teachers) spend on the creation of the portfolios raising questions about the fairness and comparability of the ratings (Koertz, et al., 1993, 1994). Second, there can be different degrees of involvement by other students in the work that is included in the portfolio (group work, class discussion and revision), and teachers can have differential involvement in the creation of student's portfolios (Herman, 1997). Both raise important questions about who is being evaluated in the performance assessment.

- ***Unanticipated consequences***. There is also a need to assess the potential negative consequences of performance-based assessments, i.e., how the measurements may be misused. This is clearly an expansion of the normal testing considerations, but if the goal of systemic reform is to spur increased student learning, the extent to which misuses of the tests can, in fact, provide a disincentive to real learning then this factor needs to be taken into consideration.

- ***Equity***. Winfield & Woodard (1994) raise important concerns that the push toward national standards, and assessment-driven school reform in particular, may bring strong negative consequences for the continued attainment of educational equity. Because minority students are concentrated in high-poverty schools, where they often receive a lower quality education, there are concerns that they will be at a serious disadvantage on the new performance tests because of their focus on higher-order skills that are often not well taught in disadvantaged schools (Herman, 1997).

- ***Establishing performance levels***. The creation of appropriate cut-points to define different levels of student proficiency can be very problematic as evidenced by the criticism lodged at the National Assessment of Education Progress (NAEP -- 1995, 1997) by the General Accounting Office (1992). Coffman (1993) makes an even more important point that gets to the heart of the problem with setting performance

levels: “Holding common standards for all pupils can only encourage a narrowing of educational experiences for most pupils, doom many to failure, and limit the development of many worthy talents.”

- ***Who should be included in the testing?*** Goals 2000 requires that states develop assessments that A...permit the participation of all students with diverse learning needs.” This raises the important question about what to do with students with either limited-English proficiency or handicapping conditions.

- ***Measuring change or gain.*** Determining what we mean by change over time is particularly difficult on these types of tests. On the one hand, the tasks that are included can vary over time raising concerns about comparability; on the other hand, increased familiarity with the assessment format may by itself create artificial test score gains. As a consequence, any attempt to evaluate students, teachers, or schools on the basis of Again scores” will entail both enormous technical and interpretation problems.

- ***Single versus multiple indicators.*** Many critics have argued that diverse indicators of performance are needed to get a more valid assessment of student achievement, e.g., using multiple teacher assessments over a period of time. “The emphasis on reform almost inevitably focuses on outcomes, particularly single measures of improvement in students’ performance. This alone is not bad, but the breathless expectation of performance is a risky criterion of success” (Chubin, 1997). Consideration must, therefore, be given to the inclusion of multiple, and intermediate, indicators of success.

- ***School and teacher accountability***. The high-stakes nature of these tests, especially when tied to systemic school reform, can raise concerns about potential abuses of the testing system. Suggested solutions include the use of a variety of quality control mechanisms such as test moderators, and selective re-scoring by external experts.

- ***Cost.*** These new types of assessment are not inexpensive. Teachers can devote substantial amounts of time to getting students ready (especially for the portfolio assessments), and the costs of scoring are far greater than for multiple choice tests -- $65 versus. $2-$5 per student (Herman, 1997).