

ADDING NEW DIMENSIONS TO CASE STUDY EVALUATIONS:
THE CASE OF EVALUATING COMPREHENSIVE REFORMS

Robert K. Yin, Ph.D.

Darnella Davis, Ed.D.

March 13, 2006

Draft Manuscript prepared for *New Directions in Evaluation*
Forthcoming Volume: Informing Federal Policies for Evaluation Methodology:
Building the evidence base for method choice in government sponsored evaluation

COSMOS Corporation
3 Bethesda Metro Center
Suite 400
Bethesda, MD 20814
301 215-9100
Facsimile 301 215-6969
ryin@cosmoscorp.com
ddavis@cosmoscorp.com

This chapter is based on research supported in part by the National Science Foundation under Grant No. REC-9970832: "Studying Statewide Systemic Reform in Seven States and One Territory;" Contract No. REC-9912173: "Cross-Site Evaluation of the Urban Systemic Program;" and Grant No. ESI-0451942: "Study of Student Achievement Trends in the Urban Systemic Program." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The Need for a Broad Array of Evaluation Methods

To benefit American society, federal, state, and local services continue to test a broad array of interventions. In turn, these varied interventions, and the need to know how well they work, demand a broad array of evaluation methods.

Only some interventions can be designed as experiments. In such cases, the experimental method provides a suitable evaluation design. If, however, a promising intervention happens not to follow experimental principles, evaluators have two choices: 1) to ignore such interventions, or 2) to develop alternative evaluation methods other than those involving experimental designs.

“Comprehensive reform” has been one such non-experimental intervention started in recent years (e.g., Schorr, 1997; Chaskin, 1999; Connell et al., 1995; and Fulbright-Anderson, Kubisch, and Connell, 1998). These reforms are still ongoing in several important public sectors:

- Systemic change in K-12 education, at state or local levels (e.g., Yin and Davis, 2006);
- Community partnerships, in relation to social and economic development as well as strengthened crime and violence prevention (e.g., Rog, 2003);
- Public health campaigns, with longstanding examples in drug prevention and the prevention of such diseases as HIV/AIDS (e.g., Scherer, 2003, 2004); and
- Mental health, including early childhood mental health (e.g., Bickman, 2005).

In nearly every situation, the reforms call for collective, coordinated, and usually concurrent actions. The actions may be wide-ranging, such as: implementing improved service practices; altering public policies; engaging citizens and the public at large in meaningful roles; providing additional resources to service organizations; and, in the case of education, implementing curricula that are aligned with prevailing tests and assessments.

The multi-faceted actions nevertheless reflect a common and single vision. As such, comprehensive reforms are aimed at changing whole systems rather than engaging in piecemeal, isolated, and sometimes conflicting initiatives. In fact, advocates of comprehensive reform espouse its promise in part *because* of the lack of prior success of the piecemeal and isolated initiatives (e.g., Schorr, 1997).¹

Thus, providing evaluative feedback about comprehensive reform continues to challenge evaluators. Some have shied away from the challenge. In so doing, however, they may be denying policymakers and citizens the opportunity to collect empirical data and information about potentially powerful instruments of improvement. Conversely, by accepting the challenge, evaluators also take risks. Insufficient evaluation methods may produce misleading data and information, resulting in more harm than good. Insufficient methods also may produce harsh peer criticisms, if not outright ostracism from the community of evaluation scholars.

Faced with this choice, our own evaluation work has continued to be based on several beliefs. First, we continue to believe that evaluators must respond to real-world conditions, not just the preferences of scholars: If public officials are trying to reform systems, evaluators ought to try to develop methods to assess those reforms. Second, we accept the treacherous path whereby the development of the needed empirical methods may require one or more generations of scholarly efforts. We hope that by persevering in forging a path, appropriate methods eventually will emerge.² Third, the path must nevertheless be laden with rigorous thinking—e.g., questioning questions, testing assumptions, and examining rival explanations as part of interpreting findings. Evaluators must constantly strive to pursue such thinking—which transcends the use of any particular evaluation method. Conversely, the sheer use of any particular evaluation method itself does not assure the presence of rigorous thinking.

¹As a side observation, the narrowness of some of these initiatives may in fact have been a direct result of the limitations imposed by using experimental principles and designs.

²Such a multi-generational perspective bears some similarity to the decades needed, starting in the early 20th century, to develop and apply clinical trial methods in medicine (e.g., Jadad, 1998).

New Demands on the Case Study Method

Practitioners of existing evaluation methods may not be convinced that comprehensive reform creates a need for newer methods. For instance, Boruch and Foley (2000) cite many studies that used experimental designs to assess community initiatives and other interventions where “sites” or organizations have been the units of allocation and analysis. Going one step further, other evaluators could argue that comprehensive reform merely requires the use of multiple, but still conventional evaluation methods. Each method could assess some aspect of the reform, and if carefully coordinated, the methods as a group could serve as the overall reform assessment.³

Our working premise, however, is that evaluating comprehensive reform may pose new demands. In fact, evaluating reform may even stretch what has been the traditional use of the case study method.

A modest starting point is to observe that evaluations of comprehensive reforms are likely to require both quantitative and qualitative evidence. For instance, the frequency and magnitude of outcomes from comprehensive reform may be best captured by quantitative results. However, these outcomes may occur in multiple domains (e.g., client outcomes and organizational outcomes), requiring separate quantitative models but then needing some qualitative technique to aggregate findings across domains. Similarly, the breadth of any comprehensive reform can create a messy array of independent variables, requiring qualitative methods to marshal the potential explanations for “how and why” the outcomes might have occurred (Shavelson and Towne, 2002). Yet, some of the potentially explanatory processes also may need to be enumerated through surveys or the tallying of archival records, thereby producing additional quantitative data. Overall, the diversity of evidence can readily partition an evaluation into multiple *parallel* but not *converging* components (e.g., Yin, 2006). Without a uniting methodology, evaluations of comprehensive reform could themselves become piecemeal.

³For instance, a National Academy of Sciences report on education research (Shavelson and Towne, 2002) recognizes that different evaluation questions lead to the need for different evaluation methods. The report notes that establishing the effectiveness of particular educational practices might be addressed by methods different from understanding how and why the practices worked. A comprehensive evaluation of such practices would therefore require the use of multiple methods. However, the report only appeals to the need for multiple methods and does not focus on the potential need for new or modified methods.

Uniting the diverse array of evidence into a coherent picture of a reform effort can be considered a chore of the case study method (e.g., Yin, 1994). The “case” is the reform effort, and the case study’s ability, if not reliance on multiple sources of evidence, enables case study evaluations to integrate the relevant evidence about reform. However, reform’s breadth and complexity may require a technical stretch even on the part of the case study method.

The classic case study focuses on single entities—an individual, organization, decision, community, and the like. When the method is used to do evaluations, the case study entities are more likely to be practices, projects, and programs (U.S. General Accountability Office, 1990).⁴ The evaluations distinguish between “context” and “phenomenon,” whereby the subject of evaluation may be considered the phenomenon of interest and the surrounding events its context. At the same time, one strength of the case study method is its ability to tolerate the real-life blurring between phenomenon and context (Yin, 2003a, p. 13). Such a methodological attribute fits reforms well.

To take but two examples, the desired systemic change in education reform goes beyond the formal K-12 system (usually a “school district”) and involves local businesses, community organizations, and other public services (e.g., Kerchner, 2005; Wong and Shen, 2001; and Yin and Davis, 2006). In education reform, the education of children and youths is not limited to what occurs in the classroom. During childhood, the home environments also assume great importance. Later on, employment opportunities during and after the formal K-12 years also may create strong incentives (or disincentives) for learning. Similarly, comprehensive community initiatives aimed at improving whole neighborhoods—or even reducing violence in these neighborhoods—may engage citizens and organizations from multiple sectors (e.g., Chaskin, 2001; and Rog, 1997). In neither the education nor community example would phenomenon and context be readily disentangled.

⁴If limited to a single practice, project, or program, a “single-case” study is the result. However, the same evaluation could cover practices at several sites that might be part of the same program. Under this circumstance, if the practice remains the unit of analysis, the result is a “multiple-case” study.

Appreciating the full complexity of comprehensive reform in fact, however, leads to a more radical revelation. In studying comprehensive reform, scholars have begun to recognize explicitly the actual confounding, not mere blurring, of “phenomenon” and “context.” Using the term “environment” rather than “context,” Supovitz and Taylor (2005, p. 208) succinctly note that:

“Social causation theory suggests that the relationship between cause and effect is moderated by the environment. The program is not the causal mechanism. Rather, the program interacts with elements of the environment to produce different effects.”

These investigators consider comprehensive reform to be “melding the distinctions between traditional conceptions of intervention and environment” (p. 208). And, as a result, the authors challenge evaluators of comprehensive reform to adopt “fundamentally different ways of thinking about the *unit of intervention*” (p. 209) [emphasis added].

Whether evaluators have successfully done so in the past, or whether they have inadvertently followed narrow and nonsystemic paths instead—possibly only evaluating specific components within a system—is still an open question (e.g., Chatterji, 2002). In this sense alone, the evaluation of comprehensive reform pushes the envelope for case study evaluations.

Our own research then may add a second revelation. Not only might the unit of intervention need to be revisited, but the reforming environment might itself have other interventions occurring at the same time as the originally targeted intervention. Thus, the cost of introducing a richer rendition of the “context” or “environment” is that a case study evaluation might have to accommodate multiple interventions occurring serendipitously and not necessarily as part of the same reform effort, even though the effort itself already may comprise many interventions.

To summarize, comprehensive reform is an excellent example of a complex initiative, bearing at least two implications for conventional evaluation designs. First, comprehensive reform potentially confounds the traditional distinction between “phenomenon” and “context,” producing a modified definition of the unit of intervention.

Second, multiple interventions, not all part of the same planned initiative but nevertheless working concurrently, may be relevant. Both conditions create new demands on evaluation methods, including the use of the case study method.

To meet these new evaluation needs, two completed studies of comprehensive reform in K-12 education (Yin and Davis, 2006; and Yin, Davis, and Schmidt, 2005) have begun to modify and extend the case study method. They serve as examples and are discussed in the remainder of this chapter.

Two Illustrations from K-12 Education

Statewide Comprehensive Reform in Eight States. The first illustration, a study of statewide systemic reform, covered seven states and one territory (eight “states”). The study accentuated the importance of tracking multiple interventions. Some of these were part of the initiative, or “phenomenon” being evaluated. However, other interventions were serendipitously occurring as part of the changing “environment.”

A federal agency, the National Science Foundation (NSF), had provided extensive financial and technical support for K-12 education reform in the eight states, over a ten-year period. The agency had made grant awards to each jurisdiction as part of a single federal program, the Statewide Systemic Initiative (SSI). Traditional evaluation designs would have considered the program to be the “phenomenon” of interest.

Using such a traditional evaluation design, an earlier evaluation had indeed tried to assess the impact of such support (Corcoran et al., 1998; Zucker et al., 1995; and Zucker et al., 1998), recognizing from the outset that federal funds are known to be but a fraction (typically around five percent) of the financial resources available to operate the K-12 systems in a state. Not surprisingly, the earlier evaluation concluded that the federal funds had only an extremely modest effect in spurring systemwide change. By limiting itself to tracing the federally-supported interventions, the evaluation also concluded that not much reform had taken place in these states (Zucker et al., 1998).

Our alternative evaluation design deliberately started from a broader view of the scope of statewide K-12 education reform. As shown in the study’s initial logic model (see Exhibit 1), the broader view incorporated such realms as “stakeholder support and

partnerships,” “state policies,” and “teacher certification requirements”—conditions that go well beyond any federally-supported intervention. Furthermore, instead of the more typical input-output arrangement of most logic models,⁵ this one posited that, over time (t_1 through t_4), changes could advance but also regress. Finally, the “height” of the figure in Exhibit 1 reflects the amount of “scale-up,” or schools and districts in a state, involved in the reforming processes.

insert Exhibit 1 about here

The design, reflected by the logic model, implicitly assumed that the relevant jurisdictions, and especially their state education agencies, might in fact also have been pursuing other significant reform interventions. For instance, some of the state agencies were implementing strong and comprehensive reform legislation, just passed by their respective state legislatures. However, because the legislation and its implementation were not part of the federally-funded activities,⁶ the earlier evaluation had downplayed such legislation and the ensuing state policies as merely part of the “context” or backdrop for the federal efforts. Thus, in comparison to the earlier evaluation, our study had deliberately expanded the “unit of intervention” to include the broader set of actions involving the state education agencies.

The study even uncovered important interventions occurring outside of the realm of the state education agencies. In particular, nearly all of the eight states had been the venue for a protracted series of state supreme court cases regarding school finance and the

⁵The conventional input-output-outcome model focuses on the “activities” (“outputs”) supported by a singular intervention, along with the resources and conditions (“inputs”) leading to the implementation of the activities and the changes (“outcomes”) emanating from the activities. The models also typically depict “contextual conditions,” but they are peripheral and external to the linear input-output-outcome flow of events. To this extent, the conventional logic models maintain a narrow “unit of intervention” that does not assume additional interventions occurring as part of the contextual conditions.

⁶In fact, federal interventions are restricted from playing any role in major actions such as the passage of state or local legislation.

equitable distribution of education resources. In most of the states, the series extended over 20 years (see Exhibit 2). The timing of these cases overlapped with the same ten-year period of time that had been the focus of support from the federal government’s program. Our findings suggested that the disposition of the state supreme court cases influenced the statewide reform of the K-12 systems (sometimes negatively) during the same period of time that was the subject of evaluation.

insert Exhibit 2 about here

Comprehensive Reform in 27 Urban School Districts. The second completed study of K-12 comprehensive education reform covered over 20 major urban (local) school districts (Yin, Davis, and Schmidt, 2005). The evaluation design was similar to but embellished the earlier study of statewide reform.

Once again, an NSF-supported federal initiative, the Urban Systemic Program (USP), had provided extended financial support for comprehensive K-12 reform, this time funding school districts with the largest central city enrollments in the country (Anderson, 2002). Each district was to use these funds to complement a broad set of reform activities, including other ongoing interventions. Policy actions by each district’s state education agency also could be relevant. Thus, the desired evaluation design again needed to potentially recognize multiple interventions occurring as both part of the “phenomenon” (the federal initiative) and the “environment” (actions ongoing elsewhere in the district, their communities, and their states).

The evaluation made two improvements over the prior study of statewide reform. The first improvement in the USP evaluation was the use of a detailed set of reform measures, assembled into a Reform Progress Instrument. The instrument calls for a series of Likert ratings covering 14 components of reform (see Appendix). Each component has multiple items, with about half of the items rated on the basis of objective data—e.g., the

number of professional development hours—and the other half on the basis of a rater’s subjective judgment.

The raters completed the instrument after reading carefully 17 of the case studies that had been completed to that point. Each case study consisted of a detailed narrative and tabular information about each district. Once the ratings were completed for all of the districts, the districts were rank-ordered according to their ratings and hence their degree of reform progress.

At the same time, the rating instrument and its measures clearly moved in a reductionist direction. Such a direction differs from the “Gestalt” or “holistic” spirit underlying comprehensive reform. Therefore, the second improvement made in the USP evaluation tried to recognize this holistic perspective by having two analysts independently and carefully read the original case studies of each district. On the basis of this reading—and reaching a holistic judgment about each district by deliberately avoiding any mental tallying of the reform components or elements—each analyst rank-ordered the districts according to the presumed extent of their reform efforts.

Comparison of the three sets of rankings, one based on the rating instrument and the other two separately produced by two analysts avoiding any reference to the rating instrument or its elements, showed a high correlation among the rankings (see Exhibit 3). The correlations were not intended to represent reliability checks. Rather, the relationships provided additional assurance regarding the rating instrument’s construct validity.

insert Exhibit 3 about here

Ongoing and Future Work

Our experiences in these two evaluations represent but part of a continuing, long-term quest. As previously suggested, improved evaluation methods are part of a journey

that can consume one or more generations of scholarship, and we hope to be contributing to this stream of work.

The work is far from complete. Specifically, our ongoing work has been trying to incorporate concurrent student achievement trends into the overall picture. Given the current policy environment that emphasizes school and system accountability in K-12 education, any complete evaluation of comprehensive reform must be able to examine its relationship with student achievement outcomes.

For reform, the requisite achievement data should cover a multiple-year time period, because reformers commonly claim that comprehensive reform may take years to accomplish. Moreover, at the national level, the reforms of interest may occur in different states, each using its own student achievement tests.

Capturing these two conditions—a multiple-year period of time for a multi-state aggregation—is a continuing challenge. One technique is to calculate “standardized slopes,” a procedure fully delineated, along with its strengths and weaknesses, in a separate paper (Yin, Schmidt, and Besag, 2006). The procedure involves a meta-analytic approach (e.g., Cohen, 1988; and Lipsey and Wilson, 1993). In this approach, the standardized slopes are considered “effect sizes” (e.g., Cohen, 1988; and Lipsey and Wilson, 1993), and the experiences in different jurisdictions (states or districts) are considered as if they were from separate empirical studies.

How best to test for any relationship between these concurrent student achievement trends and the progress, if any, toward comprehensive reform represents some of our ongoing work. This work in K-12 education also builds on earlier work in another realm, community partnerships to prevent substance abuse (Yin and Kaftarian, 1997; and Yin and Ware, 2000).

Summary

Attempts to improve communities or public service systems across the country have involved a broad array of interventions which, in turn, may create the need for a diverse array of evaluation methods. “Comprehensive reform” has become one type of

increasingly popular initiative, representing an attempt to improve whole systems, not just to change single organizations much less to implement isolated practices.

This chapter has suggested that comprehensive reform presents two conditions that stretch current evaluation designs, including the traditional use of the case study method. The first is the deliberate confounding of “phenomenon” and “context.” The second is the likelihood of multiple interventions ongoing simultaneously but under different auspices. Both conditions need to be taken into account in any acceptable evaluation of comprehensive reform. The latter part of this chapter therefore presented two illustrative studies, both dealing with comprehensive reform in K-12 public education, that incorporated these conditions by expanding the conventional use of the case study method.

REFERENCES

- Anderson, Bernice, "Evaluating Systemic Reform: Evaluation Needs, Practices, and Challenges," in James W. Altschuld and David D. Kumar (eds.), *Evaluation of Science and Technology Education at the Dawn of a New Millennium*, Kluwer Academic/Plenum Publishers, New York, NY, 2002, pp. 49-80.
- Bickman, L., and S. Mulvaney, "Large Scale Evaluations of Children's Mental Health Services: The Ft. Bragg and Stark County Studies," in R. Steele and M. Roberts (eds.) *Handbook of Mental Health Services for Children, Adolescents, and Families*, Kluwer Academic/Plenum Publishers, 2005, pp. 371-386.
- Boruch, R., and E. Foley, "The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation and Analysis in Randomized Trials," in L. Bickman (ed.), *Validity and Experimentation: Donald Campbell's Legacy*, Sage Publications, Thousand Oaks, CA, 2000, 1:193-238.
- Campbell, Donald T., *Assessing the Impact of Planned Social Change*, Paper #8, Occasional Paper Series, The Public Affairs Center, Dartmouth College, Hanover, NH, December 1976.
- Chaskin, Robert, "Building Community Capacity, A Definitional Framework and Case Studies from a Comprehensive Community Initiative," *Urban Affairs Review*, January 2001, 36(3): 291-323.
- Chaskin, Robert, *Defining Community Capacity: A Framework and Implications from a Comprehensive Community Initiative*, Chapin Hall, Chicago, IL, 1999.

Chatterji, Madhabi, "Models and Methods for Examining Standards-Based Reforms and Accountability Initiatives: Have the Tools of Inquiry Answered Pressing Questions on Improving Schools?," in *Review of Educational Research*, Fall 2002, 72(3):345-386.

Connell, James P., Anne C. Kubisch, Lisbeth B. Schorr, and Carol H. Weiss (eds.), *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*, The Aspen Institute, New York, NY, 1995.

Cohen, J., *Statistical Power Analysis in the Behavioral Sciences*, Lawrence Erlbaum, Hillsdale, NJ, 1988.

Corcoran, T. B., P. M. Shields, and A. A. Zucker, *The SSIs and Professional Development for Teachers*, SRI International, Menlo Park, CA, 1998.

Fulbright-Anderson, Karen, Anne C. Kubisch, and James P. Connell, (eds.), *New Approaches to Evaluating Community Initiatives, Volume 2, Theory, Measurement, and Analysis*, The Aspen Institute, Washington, DC, 1998.

Jadad, Alejandro, *Randomized Controlled Trials*, BMJ Books, London, 1998.

Kerchner, Charles, "Educational Reform," *The Claremont Letter*, Issue 1, Volume 1, 2005.

Lipsey, M. W., and D. B. Wilson, "The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-analysis," in *American Psychologist*, 1993, 48 (12): 1181-1209.

Rog, D. J., N. Boback, H. Barton-Villagrana, P. Marrone-Bennett, J. Cardwell, J. Hawdon, et al; "*Sustaining collaboratives: a cross-site analysis of The National Funding*

Collaborative on Violence Prevention,” Center for Mental Health Policy, Vanderbilt Institute for Public Policy, 2003, pp. 249-261.

Rog, D. J. and M. Gutman, “Homeless Families Program: A Summary of Key Findings,” in S. L. Isaacs and J. R. Knickman (eds.) *To Improve Health Care: The Robert W. Johnson Foundation Anthology*, San Francisco: Jossey-Bass Publishers, 1997.

Scherer, Jennifer, *A Guide to Collecting Data on Performance Indicators for the Centers for Disease Control and Prevention’s Prevention Research Center*, report to the Centers for Disease Control and Prevention, COSMOS Corporation, Bethesda, MD, February 2004.

Scherer, Jennifer, *A National Evaluation Plan for the Centers for Disease Control and Prevention’s Prevention Research Centers (PRCs): Final Report*, report to the Centers for Disease Control and Prevention, COSMOS Corporation, Bethesda, MD, December 2003.

Schorr, Lisbeth B., *Common Purpose, Strengthening Families and Neighborhoods to Rebuild America*, Anchor Books, New York, NY, 1997.

Shavelson, Richard J., and Lisa Towne (eds.), *Scientific Research in Education*, National Academy Press, Washington, DC, 2002.

Singer, Alan, and Michael Pezone, “Education for Social Change: From Theory to Practice,” in *Workplace*, 5.2, July 2003.

Supovitz, Jonathon A., and Brooke Snyder Taylor, “Systemic Education Evaluation: Evaluating the Impact of Systemwide Reform in Education,” in *American Journal of Evaluation*, June 2005, 26(2):204-230.

U. S. General Accounting Office, *Case Study Evaluations*, Program Evaluation and Methodology Division, Washington, D.C., November 1990.

Wong, Kenneth K., and Francis X. Shen, "Rethinking the Fiscal Role of the States in Public Education," in *Government Finance Review*, October 2001,17:8-13.

Yin, Robert K., *Application of Case Study Research*, Second edition Sage Publications, Thousand Oaks, CA, 2003, Second edition.

Yin, Robert K., *Case Study Research: Design and Methods*, Third edition Sage Publications, Thousand Oaks, CA, 2002.

Yin, Robert K., "Mixed Methods Research: Are the Methods Genuinely Integrated or Merely Parallel?" *Research in the Schools*, Spring 2006, 13:41-47..

Yin, Robert K., "Rival Explanations as an Alternative to Reforms as 'Experiment'," in Leonard Bickman (ed.), *Validity and Social Experimentations: Donald Campbell's Legacy*, Sage Publication, Thousand Oaks, CA, 2000.

Yin, Robert K., Darnella Davis, and R. James Schmidt, *The Cross-Site Evaluation of the Urban Systemic Program, Final Report, Strategies and Trends in Urban Education Reform*, submitted to the National Science Foundation, COSMOS Corporation, Bethesda, MD, 2005.

Yin, Robert K., and Darnella Davis, "State-Level Education Reform: Putting all the Pieces Together," in *Systemwide Efforts to Improve Student Achievement*, Kenneth Wong and Stacey Rutledge (eds.), Information Age Publishing Inc., Greenwich, CT, 2006, pp. 1-33.

Yin, Robert K., and Shakeh J. Kaftarian, "Introduction: Challenges of community-based program outcome evaluations," in *Evaluation and Program Planning*, August 1997, 20(3), pp. 293-297.

Yin, Robert K., and Angela J. Ware, "Using Outcome Data to Evaluate Community Drug Prevention Initiatives: Pushing the State-of-the-Art," *Journal of Community Psychology*, May 2000, 28:323-338.

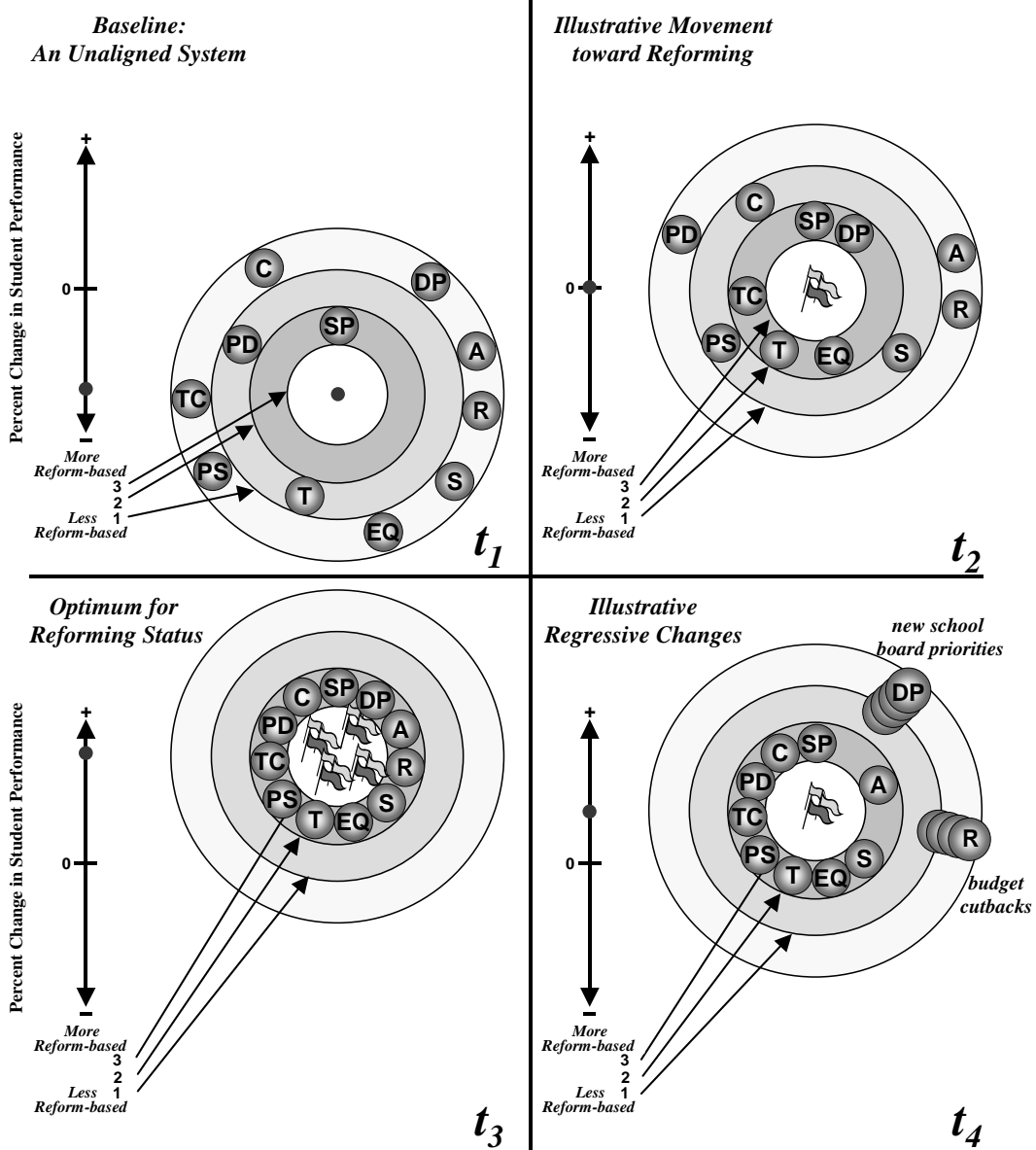
Yin, Robert K., R. James Schmidt, and Frank Besag, "Aggregating Student Achievement Trends Across States with Different Tests: Using Standardized Slopes as Effect Sizes," *Peabody Journal of Education*, 2006, 81(2):47-61.

Zucker, A. A., P. M. Shields, N. E. Adelman, T. B. Corcoran, and M. E. Goertz, *A Report on the Evaluation of the National Science Foundation's Statewide Systemic Initiatives Program*, SRI International, Menlo Park, CA, 1998.

Zucker, A. A., P. M. Shields, N. E. Adelman, and J. Powell, *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: Second Year Report. Part 1: Cross-Cutting Themes*, SRI International, Menlo Park, CA, 1995.

Exhibit 1

STATES OF A REFORMING SYSTEM



Key: A = Assessment; C = Standards-Based Curriculum; DP = District Policies; EQ = Equity; PS = Preservice Requirements; PD = Professional Development; R = Resource Convergence; S = Stakeholder Support and Partnerships; SP = State Policies; T = Technology; TC = Teacher Certification Requirements;

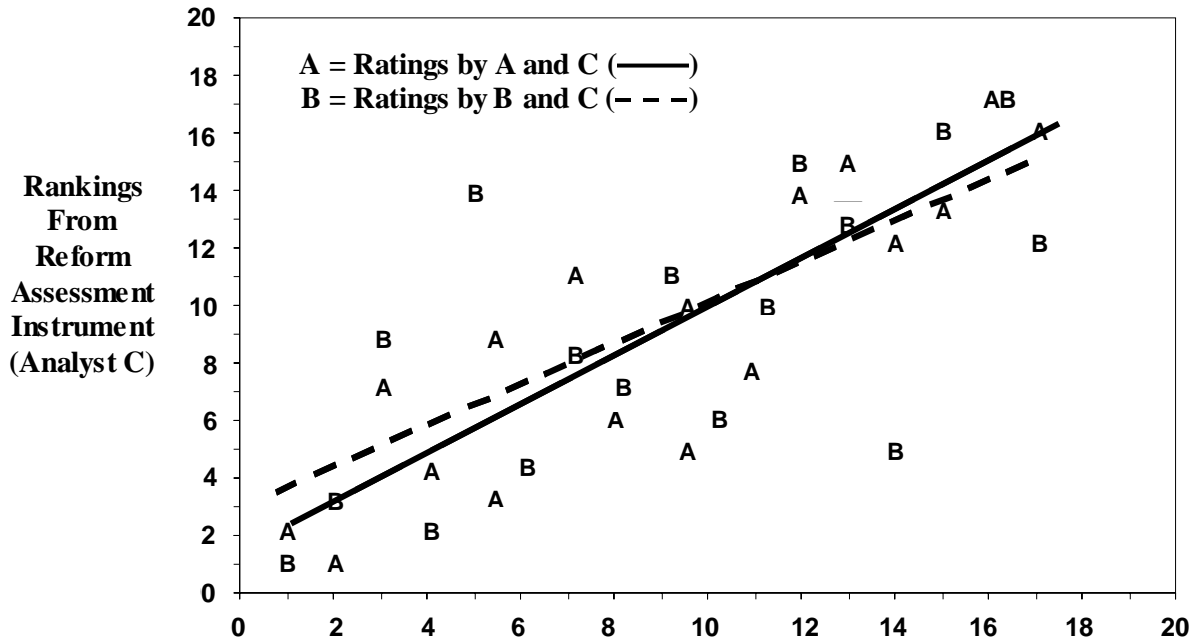
Scale up in Schools and Classrooms = Few or none ● Some Large majority

Exhibit 2

SCHOOL FINANCE COURT RULINGS AND RELATED STATE LEGISLATION, FOR 8 STUDY STATES

State	State Supreme Court Case(s)	Year of Court Ruling	General Nature of Ruling	Subsequent Legislation, if any, and Year of Legislation
CT	<i>Horton v. Meskill</i>	1977	Method for funding education declared unconstitutional	Installation of Guaranteed Tax Base formula and State Equalization Grants (1979)
	<i>Gallulo v. Waterbury</i>	1978	Districts cannot rely solely on state's general funds	Adoption of Minimum Expenditure Requirement
	<i>Horton v. Meskill</i>	1984	Court establishes three-part test for future challenges to state's funding mechanism.	Educational Enhancement Act of 1986
	<i>Sheff v. O'Neill</i>	1996	Hartford's students not provided equal educational opportunity under the state constitution (students are racially, ethnically, and economically isolated, a result of the state's creation of local school districts)	An Act Enhancing Educational Choices (1997)
	<i>Johnson v. Rowland</i>	current	Case filed in 1998, still pending. Eleven towns charge that per pupil expenditures have not kept up with inflation and that state funds have actually decreased since 1988.	
LA	<i>Chalet v. State; Minimum Foundation Commission v. State</i>	1998	Suits dismissed, remanded, and then dismissed for a second time	NONE
MA	<i>McDuffy v. Robertson</i>	1993	State fails to provide an adequate education for all its children	Education Reform Act of 1993
NJ	<i>Robinson v. Cahill, 1973, 1975, and 1976</i>	1973-76	State financing violates state constitutional mandate for a 'thorough and efficient education'	Public School Educational Act (PSEA) of 1975 and other subsequent legislation
	<i>Abbott v. Burke, 1985, 1988, 1990, 1994, 1997, and 1998</i>	1985-97	System of financing and PSEA of 1975 found unconstitutional as applied to poorer urban districts; later decision finds QEA to be unconstitutional	Quality Education Act (QEA I) of 1990; Quality Education Act (QEA II) of 1991; Comprehensive Improvement and Financing Act (1996); others
PR	NONE			
SC	<i>Richland County v. Campbell</i>	1988	Court supports lower court and dismisses equity suit challenging constitutionality of public school funding	NONE
	<i>Abbeville County School District v. State</i>	1999	Court overrules lower court, upholding claim by state's rural and poor districts that financing system violates state constitution; case remanded to trial that is still scheduled to take place	
TX	<i>Edgewood v. Kirby I</i>	1989	State's education funding scheme, created by Educational Opportunity Act of 1984 (HB 72), violates state constitution	SB1 Adjusted funding except for wealthiest districts
	<i>Edgewood v. Kirby II</i>	1991	Original remedy to 1989 ruling insufficient	SB 351 Created 188 county education districts
	<i>Edgewood v. Kirby III</i>	1992	Tax system ruled unconstitutional	
	<i>Edgewood v. Kirby IV</i>	1995	Linked equity of funding with adequacy of education	
	<i>Edgewood v. Kirby V</i>	1998	District court dismissed case because of legislative changes	SB7 Created current recapture system
VT	<i>Brigham v. Board of Education</i>	1997	School financing inequitable, and spending levels to be equalized	Equal Education Opportunity Act of 1997

Exhibit 3
COMPARISON OF RANKINGS FOR 17 DISTRICTS,
BASED ON THREE INDEPENDENT ASSESSMENTS OF
REFORM PROGRESS



Rankings From Two Analysts (A + B)
Making Independent, Holistic Judgments

Strength of Relationship

Rater	Coefficient	N	P
A + C	0.872393	17	<0.01
B + C	0.67402	17	<0.01

Appendix

SCHOOL DISTRICT A*

REFORM COMPONENT	RATING	MATHEMATICS AND SCIENCE	
A1 Standards-based Curriculum	12	MATHEMATICS	SCIENCE
		District uses state recommended textbooks and materials such as <i>Math Their Way</i> , <i>Connected Math</i> , <i>Sharon Wells Math</i> (all adopted under CPMSA), reported to be aligned with state framework and assessments. In 2004-05, District adopted <i>Math Investigations</i> districtwide and also introduced <i>Connected Mathematics</i> at all middle schools. That year, the number of AP enrollments continued to decline as more students took three and more years of mathematics or science although the district pays for first-time AP exams.	Adopted and implemented <i>FOSS</i> for elementary science in 1999-2000. Uses state text as resource to district's policy of 40% lab time drawing on <i>FOSS</i> , <i>TEXTEAMS</i> , and Dana Center's <i>Science Toolkit</i> . Adopted <i>CPO Cambridge</i> science curriculum for secondary level in 2002; is on state approved list and vertically aligned with elementary and middle school science curriculum. Revised most secondary level documents in 2001-02 to align them vertically with <i>FOSS</i> ; math curriculum documents revised to ensure alignment with state standards. <i>Cambridge Physics</i> selected for secondary science in 2002-03 and aligned with elementary, middle school curriculum.
A2 Assessment	12	In 2003 students began taking TAKS which replaced TAAS as the statewide assessment; schools were accountable for TAKS in 2004. TAKS will serve as gateway test for promotion in grades 3 (in reading), 5 and 8 (reading and math), 11 (as exit exam for high school diploma). TAKS will be offered in science in grades 5, 8, and 11. The previously state required end-of-course exams in Algebra I and Biology have been folded into the TAKS. SAT-9 open-ended math and science tests being administered to samples of second graders since May, 2000 as part of a longitudinal study. District uses benchmark tests designed to assess students on TEKS standards expected to be addressed during a specific time frame for insights into student instructional needs. Science benchmark tests given in 3rd, 5th, and 8th grades and for biology, integrated physics, and chemistry; math benchmark tests given in grades 3-8 and for Algebra I.	
A3 Professional Development	12	District policy requires time and financial resources to support staff development based on teacher and administrator needs. District requires teachers to take 18 hours of PD annually and to attend all "waiver days." In addition, new teachers must attend 2 days of orientation PD. Average of 36 hours for secondary math teachers and 27 hours for science with more than 48 percent receiving some training in 2004-05. District tracks electronically all PD offered. District received 4 "waiver days" (2 district- and 2 site-level) from state, in 2004-05. Administrators required to take 200 hours (100 of instructional-oriented and 100 of leadership and administrative professional development). PD during 2001-02 focused on the use of <i>FOSS</i> kits for elementary level teachers and on vertical alignment across subject areas and "teaching to the TAKS" for secondary level teachers. In 2001-02 USP sponsored training in how to complete observations of inquiry-based classrooms for elementary level principals and in 2002-2003 in how to evaluate teachers in inquiry-based settings for assistant principals as well as the statewide teacher evaluation system focused on standards-based instruction. By 2004-05, the focus was on site-based PD supporting curriculum implementation for, e.g., <i>FOSS</i> , <i>TEXTEAMS</i> , exemplar labs, <i>Connected Math</i> , <i>Math Investigations</i> .	
A4 State Policies	3	Eliminated all courses below Algebra I during 1997-98. Mandated state commissioner to establish master math teacher grant program for 2003-2004; similar program being designed for technology. State commissioner also required to develop math training materials and teacher training resources to help math teachers develop expertise in math pedagogy.	

Continued

REFORM COMPONENT		RATING	MATHEMATICS AND SCIENCE
A5	District Policies	6	Eliminated all courses below Algebra I during 1995-96. Course requirements exceed those of state high school diploma; beginning with 2001 freshmen, 4 years of math and 3 years of science including biology and chemistry or physics. USP worked with district to get more science instruction at the elementary level; 15% of instructional day must be devoted to standards-based and inquiry-centered science instruction. Math and science have been semi-departmentalized for the first time for grades 4 and 5.
A6	Preservice and Teacher Certification	3	The local IHE working to institutionalize pre-requisites, expectations, and requirements of teachers. The local IHE developed 2 new science courses required of elementary education students, science faculty using FOSS kits as instructional tools for undergraduate and graduate courses and <i>Connected Math</i> with science teachers. State requirements for certification apply; changed in 1999 requiring more science and math content courses for teachers through 8th grade; math and science courses for preservice teachers must be aligned with standards. Alternative teacher certification available for candidates who complete a BA, maintain a 2.5 GPA, take 24 hours of alternative certification coursework, and pass the relevant content portion of the state assessment in addition to completing a 1-year internship.
A7	Teacher Recertification	3	All teachers required to take 180 hours of PD within five years to retain certification since 1999. Prior requirement was 90 hours. PD is based on district or school identified needs.
A8	Resource Convergence	3	District uses Title I, II, Eisenhower, Bilingual Education, Texas Compensatory Education funds, state, and federal technology grants, targeted math and science related external funds through alliance with the local IHE. Funds obtained through alliance with the local IHE and other initiatives by the state expected to continue after USP funding ends. USP has little influence on how federal, state, and local funds are used.
A9	Technology	2.5	Technology integration is part of District's strategic plan with goals set for technology use in all secondary classrooms.
A10	Stakeholder Support	3	Alliance between the local IHE and American Physicists Association to develop more inquiry-based instructional activities and a teacher-scientist alliance. District has partnerships with multiple community organizations aimed to improve teaching and learning of science and math. The local IHE and USP funding training program for 6th grade teachers seeking math certification. The local IHE and USP developed jointly a course for elementary level education majors in Physics (Physical Science for Teachers) offered by the department of Physics that by 2003-04 was a requirement for potential elementary teachers.
A11	Equity	6	Disaggregated results of end-of-course examinations for 2001-2002 in Algebra I and Biology demonstrated Hispanic and economically disadvantaged students outperformed similar groups at regional and state levels with no gaps when compared to all test takers.
A12	Scale-Up and Implementation Strategy and Progress	6	District implemented all districtwide activities during the first USP year. Greater emphasis on science at elementary and secondary levels in 2001-02 and 2002-03; activities in math were considered "maintenance activities" to build upon and support work done under CPMSA. 16 mentor teachers provided onsite assistance to teachers; each mentor supported 4 schools and worked primarily with a different grade level each year. Support for other teachers came through districtwide PD and in response to individual requests. Mentors provided special attention to 5th grade teachers in 2002-03 because their students would take the science TAKS; they spend additional time with novice teachers.
B1	Organizational Positioning and Leadership	3	USP's project director coordinates district science and mathematics activities including professional development, technical assistance, curriculum selection and implementation, and assists in interpreting student and teacher performance data to improve instruction. The project director reports to the asst. superintendent for curriculum and instruction, who reports to the superintendent.
B2	Absence of Events Potentially Disruptive to Standards-Based Reform	3	Statewide accountability system is central to all district efforts of reform. CPMSA award between 1994 and 1996 allowed district to implement <i>Math Their Way</i> , <i>Sharon Wells</i> , and state adopted curricula; CPMSA did not include science. Most students who enter university are graduates of the district's schools; 80% of the district's teachers are graduates of the local IHE. District accountability practices fully comply with NCLB requirements and, after an appeal, for 2003-04 the district met AYP.
Total		78	

*A rating scale was defined for each of the 14 components (with certain key components weighing more heavily than others (see below for individual weights). Each site was then ranked according to the total of their component scores and placed in one of four different ranges—roughly coinciding with a sense of “high

(above 69),” medium (between 60 and 78),” “moderate (between 54 and 59),” and “low (below 59).” This example is from Group 1, and is classified as “high” reforming.

Rating Key

- | | | |
|--|--|--|
| A1 Standards-Based Curriculum = 12 | A6 Preservice and Teacher Certification = 3 | A11 Equity = 6 |
| A2 Aligned Assessment = 12 | A7 Teacher Re-Certification = 3 | A12 Scale-Up and Implementation Strategy and Progress = 6 |
| A3 Professional Development = 12 | A8 Resource Convergence = 3 | B1 Organizational Positioning of USP Leadership = 3 |
| A4 Supportive State Policies = 3 | A9 Technology = 3 | B2 Absence of Events Potentially Disruptive to Standards-Based Reform = 3 |
| A5 Supportive District Policies = 6 | A10 Stakeholder Support and Partnership = 3 | |