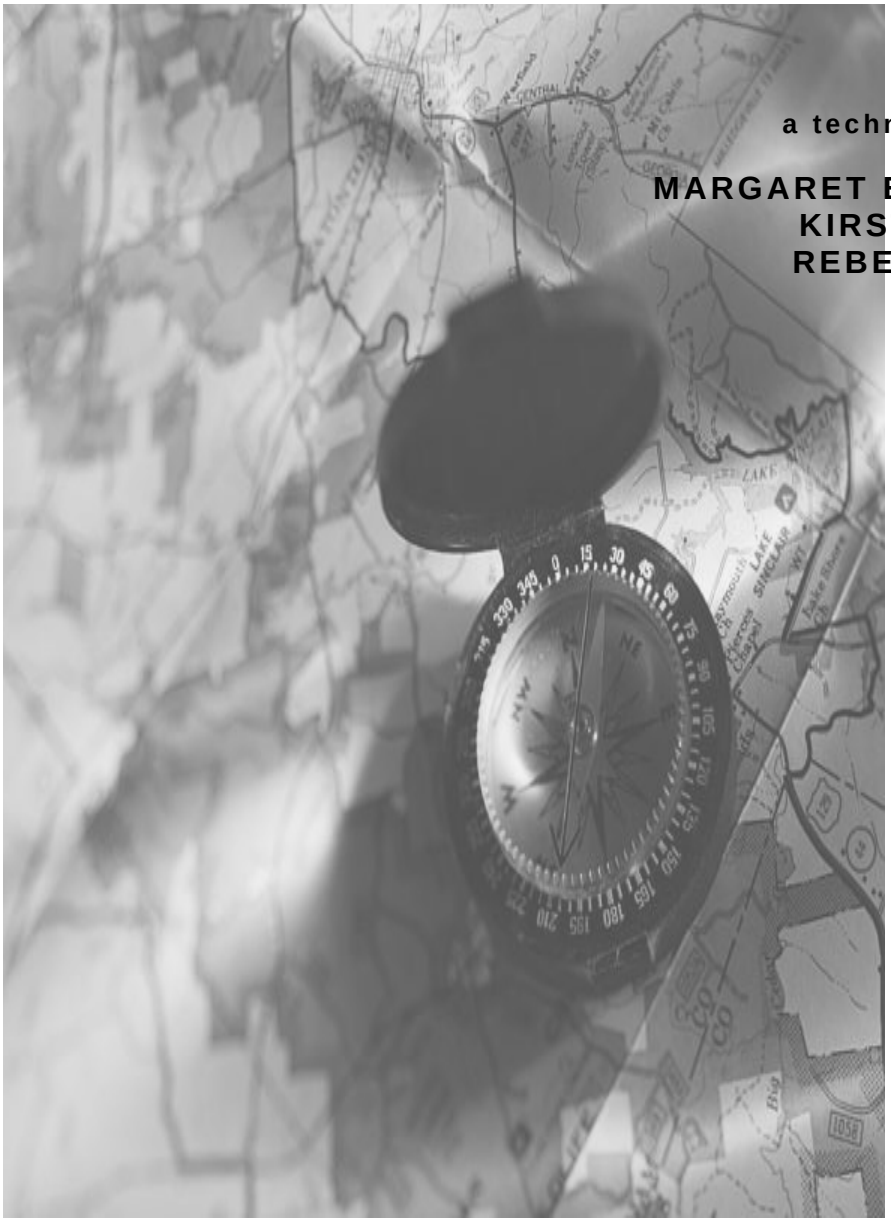
PROGRAM EVALUATION CAPACITY IN STATE DEPARTMENTS OF EDUCATION



a technical report by

MARGARET E. RAYMOND
KIRSTY BORTNIK
REBECCA GOULD

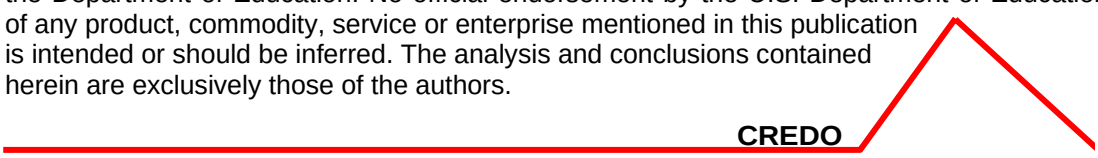
CREDO

CENTER FOR RESEARCH ON EDUCATION OUTCOMES
Hoover Institution Stanford University Stanford, CA

Center for Research on Education Outcomes (CREDO)
Hoover Institution
Stanford University
Stanford, CA
http://credo.stanford.edu

April 2004

CREDO gratefully acknowledges the support of the U.S. Department of Education for this project. The views expressed herein do not necessarily represent the positions or policies of the Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this publication is intended or should be inferred. The analysis and conclusions contained herein are exclusively those of the authors.



Contents

Tables

Preface and Acknowledgments

The support and assistance of many people were essential to the completion of this research and final report. We would like to express appreciation for those efforts here.

Alan Ginsburg and David Goodwin of the United States Department of Education provided encouragement, guidance, and support throughout the project. In addition, they willingly shared their unparalleled experience in the workings of federal and state evaluation practice. Without their insights and suggestions, the project would have been less focused.

The report would have been impossible without the cooperation of the many individuals in the state education departments who cooperated with our requests for information. The effort required to provide the level of detail needed for this research was considerable, especially in the light of hindsight. Many respondents turned our inquiries into labors of love. Moreover, most of the respondents were attuned to the likely conclusions of our research and still were candid during our interviews. Our understanding of the current program evaluation abilities of state education departments is founded on the efforts of our confidential respondents, and we are grateful.

We would like to thank our independent reviewer Carol Weiss for her diligent review and helpful suggestions on our preliminary drafts. Many of her questions caused us to delve back into the data to gain a deeper understanding of state behavior. Her contributions were important and welcome.

Program support for this project came from the U.S. Department of Education, the Smith Richardson Foundation and the Packard Humanities Institute. Their interest in the questions addressed by this research provided a strong base from which to launch the project. We thank them for their generous support.

Project Staff

Margaret E. Raymond, Project Director
Kirsty E. Bortnik
Rebecca J. Gould
Kate M. Chauncy
Stephen H. Fletcher

Administrative Staff

Claudia Traver
Cerena Sweetland-Gil
Sarah Kinkel

Executive Summary

Introduction This report presents research findings on the current capacity of departments of education in the fifty states to perform rigorous evaluations of programs in public education. A joint project by CREDO, a policy research group at the Hoover Institution of Stanford University, and the Outcomes Research Institute of Stanford, California, was completed for the United States Department of Education in 2001 – 2002. This report on program evaluation capacity in the states is timely. Increased attention to empirical evidence about program performance is evident in the numerous references to "research based" policies in the *No Child Left Behind* legislation. In its sponsorship of the What Works Clearinghouse, the United States Department of Education has also reinforced the importance of high quality program evaluations as the basis for policy decisions.

Methods A survey was administered to the departments of education in the fifty states in 2001 and 2002. Forty-nine out of fifty states agreed to take part in the survey, though the degree of participation varied. The survey investigated five areas: the current capacity to field evaluation studies, the prevalence of evaluation activity in the previous five years, the rigor of the research designs employed to evaluate programs, the mechanisms used by education departments to share study results and incorporate findings into future decisions, and perceived deficiencies or opportunities for future action. To allay states' concerns about negative repercussions from their participation, the researchers agreed to present the findings in aggregate and not attribute specific examples to the states concerned.

Evaluation Activity The findings show that most states infrequently evaluate any their programs, if at all. Nine states (18 percent) did no evaluations over the five years. Another nine states did less than one evaluation per year. In 18 states (36 percent) one to two evaluations per year were reported. Only in 12 states (24 percent of all states) did the level of evaluation activity produce eleven or more evaluations in the five-year period of study, for an average that exceeds two studies per year. Thus, about a third of states do practically none, another third does a little, and a third does a noticeable number of evaluation studies.

Review of Identified Evaluations Details on 273 evaluations were obtained from the states and were used to examine the context for the evaluations, the use of resources for the evaluation and the approaches employed in their execution. More than half of the evaluations (55 percent) were initiated by outside parties – by state legislators or by federal programs operating through the state departments of education – as a condition of funding. The remainder originated with the departments of education, where program staff sought program performance evidence roughly three times as often as the agency executives. When evaluations were done, only a third had direct involvement by education department staff – two thirds were assigned to outside consultants or academic researchers.

Given the reliance on outside expertise, the results on research designs and methods were especially worrisome. Many program evaluations contained a mix of approaches: about a third were formative evaluations, one half included process evaluation designs,

and slightly over half were designed as impact evaluations. The large number of formative and process designs helps but does not fully explain the predominant reliance on trend analysis (used on 25 percent of evaluations), focus groups (used in 26 percent of program evaluations) and satisfaction surveys, which were employed in over 40 percent of the evaluations. Random assignment or quasi-experimental designs – the only methods that control for program participation and other countervailing factors and therefore produce reliable results – were found in only 21 percent of the evaluation studies. The results are even sharper when the pool of studies is restricted to those studies that are intended to examine impacts of programs: less than 10 percent of all the studies purporting to be impact evaluations used random assignment or quasi-experimental designs.

Departmental Management of the Evaluation Practice In addition to designing and executing program evaluations, it is necessary for states to generate the studies, provide oversight and review of the deliverables, vet the results and put the results to use. We refer to these activities as evaluation management, and they can affect in important ways the number, quality and ultimate utility of the evaluations that are done. The findings show that in states that organize a separate Research and Evaluation unit within the department, the level of staffing devoted to evaluation activities is significantly higher than in states that handle their internal evaluation practice by drawing on staff from other units in the department. However, staffing levels were not fod unto significantly affect the use of outside evaluation professionals. In addition, the

presence of a research and evaluation unit was not significant in determining the number of studies that were produced.

The findings on evaluation management activities show that few states follow regular procedures, even when policies exist. Often states patch together ad-hoc arrangements, with little attention to on-going quality assurance. Dissemination of results does occur, but at the discretion of the departments. This finding raises the risk that other parties involved in building education policies are not fully informed about the performance of the programs that are evaluated, which are themselves a small subset of the number of programs operating under department auspices. There are no formal means to consider the evaluation findings in future decisions.

Visions of Future Development The representatives of the state education departments who responded to the survey were aware of the shortcomings in their organizations. They offered over 100 suggestions for changes in their departments that would increase the role of evidence on program performance. The suggestions sorted into general categories that revealed that many states shared the same development interests. The two most frequently mentioned areas for improvement are related: more resources for evaluation (mentioned by over half the states) are directly related to better institutional support for program evaluations in current operations (mentioned by 40 percent of the states.) More highly trained staff was also mentioned by more than a third of the states. These suggestions indicate that there may be a role for the US Department of Education or other external organization to foster growth in this area of

state education departments, or for states to achieve some scale economies through collaboration to address some of these suggestions.

Discussion This survey was an important first effort to examine the current capacities in state education departments to critically assess the performance of their programs. The first objective – to adequately describe the state of the states – was fully accomplished, with dramatic and surprising evidence that states are not giving as much attention to program evaluations as one might think. The second objective of the project – to explain what causes the observed differences between states in their level of effort – unfortunately did not produce definitive answers.

Perhaps the lack of explanatory power across the states is less important than the normative question that the overall findings raise. What should states be doing going forward? Giving the benefit of the doubt, it could be that to this point state education officials were unaware of the deficiencies in their departments. Since this report will be distributed to each state education department as a courtesy for participating in the study, they will now have a clearer picture of the state of affairs. The respondents from the state education department were themselves able to identify steps that could make dramatic strides in the capacity of states to review the effectiveness of the programs they sponsor. Whether states move to adopt such changes will reveal a lot about the underlying motivations and cultures in the departments. Their individual responses to these findings present an excellent opportunity for further study.

I. Introduction

As a nation we are desperate to improve failing schools and help students become fully prepared to take their places in society. With decades of flat performance and escalating costs, the public has escalated calls for educators to take decisive action to improve the situation. Calls for better instructional practices, better teacher preparation, and better leadership arise from all corners of the country.

It would be natural to assume that policy makers and educators alike insist that programs prove their worth using high standards of evidence and to use that information when choosing programs to support. But that is not the case. By not demanding better information about education program performance, we are missing out on a faster course of correcting the problems of low student achievement in the United States.

That assumption is certainly borne out in the language of the new federal legislation, the No Child Left Behind Act of 2001. With its passage, the United States Congress, President George W. Bush, the U.S. Department of Education, and much of the American public declared a unified intent to pursue research-based policies in elementary and secondary public education.[1] The legislation itself calls for research-based programs or policies over 100 times. Educators and researchers alike are aware of the low quality of much of the existing education research and are determined to improve the stock of future

[1] Public Law 107-110 *No Child Left Behind Act of 2001*, January 8, 2002.

work.[2] The shift recognizes that the last 30 years of education efforts have produced little results in improving the education outcomes of America's children. With such a strong focus on scientifically grounded policy development, an important research question becomes: "How prepared are the relevant decision makers in education to identify, create or use research in the course of their work?"

Since the primary locus of education policy rests with the states, a reasonable first cut at the issue concentrates on state education departments, the executive state agencies charged with managing the education offered to the state's students. The state education departments are keystones in the new education architecture, so their capacities for scientific research and program evaluation are of critical importance if the new policy directions are to be successful. Thus, the prime motivation for this study was to ascertain the degree to which states are capable of rigorous analysis of the performance of the elementary and secondary education programs they oversee. This report presents the results of a survey funded by the United States Department of Education and conducted by CREDO, an education policy research group at the Hoover Institution of Stanford University and Outcomes Research Institute.

The findings show that state education departments are currently ill equipped to provide the rigorous assessment of their programs that will be necessary in the future policy landscape. While departments were found to vary in the breadth of their evaluation

[2] Coalition for Evidence-Based Policy. (November 2002) *Bringing Evidence-Driven Progress to Education: A Recommended Strategy for the U.S. Department of Education.*

practice, from doing nothing to doing a fair amount, the types of studies that they undertake are unlikely to help discern good programs from bad. State education departments routinely devote few staff resources to the research and evaluation activities that go on, relying largely on outside experts to conduct the work. In fact, a substantial number of states have no formal research and evaluation division. Despite considerable expenditures for program evaluations, critical review of evaluation designs and results is the exception not the rule. As well, a range of other organizational and administrative practices that could increase the attention to evidence was found to be missing or haphazard in a large number of states.

The findings reveal that states have few mechanisms to create sufficiently high quality program evaluations. The result is that policy makers do not have the necessary confidence to rely on them in making decisions. In practical terms, the findings suggest that other non-empirical factors are the sole basis for current decisions, and there is little ability to distinguish good programs from bad. The implications for decision-making are clear.

The Case for Evidence

The need for solid evidence on program performance is unequivocal. With an annual public outlay for K-12 education of $400 Billion, policy makers are stewards of the public purse when making investments in education. As such, they have an obligation to see if the choices they make are effective. In contrast the annual national budget for

evaluating education programs is far less than $1 billion. The prospect of redirecting just a small share of the total education budget to more effective or efficient uses would crate savings each year far in excess of the costs of evaluation. And these savings would accrue over time.

States have taken steps to adopt accountability programs to gauge the performance of schools. These programs provide—for the first time in many states—a consistent measure of results. The implementation of NCLB ensures the measures will be comprehensive by covering students in grades 3-8, though the choice of measure may differ from state to state. While the designs for these programs will no doubt be refined, these initiatives are noteworthy in two ways: 1) they provide a common basis of assessment, and 2) they use outcome measures as a solid foundation on which to base rigorous examination of the effectiveness of education policy and programs. This groundwork is monumentally valuable to future efforts to discern which programs work and which do not.

An ability to prove in program effectiveness is critical. Public education today is bursting with innovations designed to improve the performance of American students. We need to know if new programs are effective for the average student, but we also need to know they work under a variety of conditions—for different types of schools or different types of students. Program designers need performance evidence to know how to improve their programs. How are we to ascertain which proposals are superior and which are inferior if we do not evaluate them on common grounds? Without the proof

that program evaluation can provide, policy makers are left to use far more subjective criteria in charting the course of education reform.

Two examples illustrate the opportunities that are lost by failing to undertake evaluation studies. The first concerns effective math practices. More than 80 Massachusetts schools have adopted the Singapore Math curriculum in an effort to improve student learning of mathematics. Singapore leads the world in mathematics performance as measured by TIMMS tests of students. Although the program shares several attributes with the approach endorsed by the National Commission on the Teaching of Mathematics, the programs differ in material ways. Whether the program is successfully implemented, whether the program operates as well in a different cultural setting and whether it produces levels of performance or gains in the United States that are equivalent to those obtained in Singapore are open questions. The circumstances are ideal for a serious effort to evaluate the effectiveness of Singapore Math in comparison to current methods. Even though Massachusetts would have to grapple with significant evaluation design issues—students are tested in non-consecutive years, making it difficult to track students longitudinally—it still would be possible to design a rigorous evaluation of the effectiveness of Singapore Math versus NCTM Math.

With such evidence, Massachusetts would be better equipped to consider expansion of the Singapore Math program to other schools. By extension, other states interested in the same issue could benefit indirectly from solid evaluation results. The U.S. Department of

Education is examining math instruction more generally and could incorporate the results of an evaluation on Singapore Math into its deliverations.

A second example concerns teacher preparation methods. Under NCLB, the Title II Teacher Quality State Grant Programs have been expanded. The former Eisenhower Program, previously devoted to the professional development of teachers in math and science, is not open to all fields. Millions of dollars are being devoted to teacher development across the country. States are adopting new approaches to teacher preparation—mentoring of new teachers, alternative certification, new ways to establish Master Teacher qualifications—often several approaches within a single state, without considering the best way to measure their effectiveness. Even if only one approach in each state were examined thoroughly, we would quickly add to the stock of knowledge about the best ways to produce high quality teachers. The final payoff would be greater coherence in the policies that were adopted.

The Challenge

Despite the current levels of effort and quality, the capacity of state education departments to undertake evaluation research is not static. It is possible to expand the role of evaluation in state education departments. One has only to consider the parallel case in human welfare agencies that oversee public assistance programs to see not only that organizational development is possible, but also that such efforts have borne fruit. Twenty five years ago, state departments of human services were in much the same

position that state departments of education now occupy: not much work on evaluation was done, and what was undertaken suffered from poor quality. With direction and leadership of the U.S. Department of Health, Education and Welfare (and later the U.S. Department of health and Human Services), states were encouraged to experiment with new approaches to public assistance under Title XIX of the Social Security Act, but only with the proviso that each experiment be accompanied by a solid evaluation of its effects. Over time, the ability of states to develop, execute, or contract for high quality evaluations increased, as did the body of empirical evidence of the inherent design flaws with the public assistance programs of the day. The resulting reform of welfare TANIF carries with it a requirement for continuing evaluation of the effects of state efforts, and many states perform these studies unassisted.

Background

The United States Education Department funded a joint project by CREDO, an education policy research group at the Hoover Institute of Stanford University and the Outcomes Research Institute to survey the fifty state education departments and analyze their existing capacity to perform program evaluations and related research. The project pursued two objectives. The first objective was to describe what is currently happening in state education departments in the area of program evaluation. The second objective was to identify key causal factors to explain the variation that was observed across states, in an effort to isolate opportunities to leverage future improvements.

This project represents a diversion from the typical reviews of research and evaluations. In the past, most reviews or meta-analyses have selected a group of studies based on a common research topic and have attempted to synthesize the findings.[3] This project complements those earlier efforts with a new sorting approach. Here, the attempt is to examine the body of work produced by the states across topics. By selecting studies by sponsor, we create the opportunity to examine the current and potential contributions of the states in developing evidence and using it in their own decisions.

At the outset it is important to clarify what is meant by program evaluation. We use the term to mean a deliberate and reflective examination of program activity based on established research principles and employing verifiable analytic tools. Delineating program evaluation in this manner distinguishes it from regular program management and monitoring which seek to describe activity within the program based on case flows or units of service to answer the question "What are we doing?" With program evaluation, the overriding questions are "How are we doing?" and "What impact are we having?" With either question, it is necessary to use some independent standard of performance to compare against the program under study.

[3] An excellent example can be found in Herman, R., Carl, B., Lampron, S., Sussman, A., Berger, A., and Innes, F. (March 2000). *What we know about comprehensive school reform models*. Report for the U.S. Department of Education, Office of the Under Secretary of Education, Planning and Evaluation Service, American Institutes for Research, Washington, DC.

The release of this report in late 2002 is timely. Earlier this year, the US Department of Education refocused its resources and organization to provide greater attention and emphasis on empirical evidence about program performance[4]. The new concentration is perhaps best exemplified by the sponsorship of the What Works Clearinghouse to vet existing research and develop syntheses of what the research demonstrates[5]. Part of the intent of the Clearinghouse is to establish uniform standards for judging the quality and rigor of education research, with, the expectation that over time more studies will meet the standards for inclusion in the syntheses. This project provides a complementary look at the production of studies – useful both as a baseline and as a diagnostic tool for future organizational development.

[4] US Department of Education (April 5, 2002) *New Directions for Program Evaluation at the U.S. Department of Education.* Press Release and Report

[5]US Department of Education (August 7, 2002) *U.S. Department of Education Awards Contract for the "What Works Clearinghouse"* Press Release and Report

II. Methods

A survey was administered to the departments of education in the fifty states in 2001 and 2002; the instrument appears in Appendix A. The survey used retrospective inquiry to identify the evaluations conducted or sponsored by the education departments over the period 1997-2001.

By keeping the inquiry concrete, we expected it would be easier to investigate the five areas of inquiry:

- The current capacity of the education department to field evaluation studies.
- The prevalence of evaluation activity in state departments of education.
- The rigor of the research designs employed by state departments to evaluate programs.
- Mechanisms to share evaluation study results and incorporate findings into future decisions.
- Perceived deficiencies or opportunities for state departments concerning program evaluation.

Completing the field portion of the project proved more challenging than originally anticipated. We found that state education departments are highly fluid environments with frequent programmatic and staff turnover. This volatility contributed in several states to a shallow institutional memory about historical events. Second, in many departments the organization is highly decentralized, and evaluation is left to the

discretion of individual program offices. With no central channel for information either up or down, the department leadership was in many cases uncertain about the use of evaluations or the results of the studies that were completed. Finally, despite cooperative efforts from most department leaders to identify the most appropriate respondent in their organizations, the survey depended on the goodwill and cooperation of individuals with other often pressing priorities; some respondents were more diligent than others in tracking down the detailed information we requested. If needed, the websites of the state education departments were accessed to obtain available information to augment the survey responses.

The survey team went to extraordinary lengths to complete the questionnaires. The survey protocol called for an initial telephone contact, a preliminary review of the project in order to secure informed consent to be interviewed, transmittal of the survey form and a combination of telephone, email and postal follow-up. In many states, the field team conducted 20 or more telephone or email contacts to obtain information needed to complete the survey.

Despite the various limitations discussed above, the research team was able to gain the participation of 49 of the 50 states. To be clear, some states provided "bare bones" responses to the survey, leaving some of the detailed questions unanswered. Thus in some states a great deal is known about the evaluation practice is known and for other departments we have learned only a little. One practical consequence is that the results often use different numbers of cases.

In conducting the field survey, it was important to address the concern of many states about the risks of exposure from participation. Particularly for those that had little or no recent experience evaluating of their programs, states did not want their participation to trigger adverse consequences. While a few respondents viewed the survey as an opportunity to see how they compared to others or as a potentially valuable tool for securing additional resources for evaluation in the future, about a third of the states' respondents were cautious about future regulatory requirements or external investigation by government or the media. Regardless of the perceptions of the survey and its consequences, most respondents looked for assurances that their state would not be singled out. Accordingly, the results are presented in aggregate; where particular examples are mentioned, they are done without identification.

A Construct for Studying States' Capacities to Conduct Program Evaluations

Determining the capacity of states to perform program evaluations presented a measurement challenge: what common elements of an evaluation practice can be identified. How best should they be put forth in an interview? What measurements should be used to accurately reflect what is happening in the states but also allow for true differences among them to emerge?

To that end, the survey solicited information on the volume of evaluations completed from 1997-2001[6]. It sought information about the internal and external resources available to assist in the evaluations that a state chooses to pursue. States were asked about the way they managed their evaluation activities and how they handled the results of completed studies. In short, they were asked to explain all aspects of their involvement in program evaluations.

Beyond the factual details of states' experience, we wanted to explore if the observed variations could be explained by obvious characteristics of the states. Did states with larger student populations behave differently than those with smaller numbers of students? Did states that reported large volumes of studies have different types or amounts of resources available for evaluations? Were states in different regions of the country more likely to follow their local peer states and have similar practices?

The information provided by states revealed considerable spread across states in their facility to perform evaluations of their programs. However, these differences do not vary significantly with any of the simple explanatory factors we studied. Instead, on almost every measure analyzed, the distribution of results yielded surprises. No neat patterns of association emerged, suggesting that factors such as having a culture of accountability or a continuous improvement ethos–things difficult to capture via limited exposures to each

[6] States were asked to report *completed* evaluations, but several included projects that were in process as well.

state – may better explain why some states have highly sophisticated evaluation practices and others have none.

Despite the lack of structural associations in the results, they nonetheless tell a fascinating story about a critical part of our ability to find solutions for the problems in American education.

III. Evaluation Activity

A threshold question for each state was whether or not they had conducted any evaluations in the previous five years. The time period was chosen to reflect a reasonable period for states to decide to undertake some study of program performance, even if it is not a regular department activity. The period also allowed for the possibility of political changes in the state or federal level that might have influenced either the expectation that evaluations would be undertaken or altered the resources that were made available for such purposes. The findings show that only a few states have a regular practice of doing program evaluations. Most states infrequently examine any of their programs, if at all.

Nine states (18 percent of all states) reported they had no evaluations done during the period of study. Since it was conceivable that these states made an affirmative decision to defer all research to specialists outside the department, the survey asked these states if they had a formal link to any external research organizations. Four of the nine indicated they had such an arrangement; three used regional labs and one was aligned to a local university. In all these cases, the research groups set the agendas for evaluation, not the states. The remaining five states neither did evaluations themselves nor made arrangements to have others do it for them.

Forty-one states reported they had some degree of evaluation activity in the past five years. In Table 1, the states are displayed by the extent of their activity.

Table 1 Evaluation Activity in the States

Number of Evaluations During 1997 – 2001	Number of States	Percent
No evaluations	9	18
1 to 5 evaluations	9	18
6 to 10 evaluations	18	36
11 to 20 evaluations	5	10
21 or more	7	14
Do evaluations- Don't know volume	2	4
Total number of States	50	100

The figures above show that, based on output, the states cluster roughly into thirds. The frequency of 1 to 5 evaluations over the five-year period translates one evaluation per year or less. When added to the states that report no evaluations, they amount to a third of states that do little or no work in this area. Roughly another third report a regular practice of evaluation averaging about 1-2 evaluations per year. The last group, amounting to only a quarter of the states, has an extensive practice of program evaluations over the past five years of study.

Of the 41 states with recent evaluation experience, we were able to secure additional information about the completed studies from 39 of them. The details about the evaluations they conducted underlie the analysis concerning the focus and rigor of recent evaluations presented in Section IV of the report. Additional detail about the way state departments of education organized evaluations and the ways they put the completed studies to use are discussed in Section V.

IV. Review of Identified Evaluations

An in-depth look at the program evaluations done under state auspices offers the basis for describing the current capacity of education departments to perform relevant and reliable program evaluations. An underlying assumption to this section is that states will seek to obtain evaluations that use appropriate tools and techniques to produce the most defensible empirical record on which to make judgments about programs. This is not to suggest that all studies should use a single approach. Nor does it imply that all studies will achieve a mandatory level of precision or quality. Instead, the assumption allows us to take the completed evaluations as approximations of the states' current but unobserved standards for what evaluations should be.

Across the 41 states known to have completed evaluations in the period 1997 to 2001, a total of 366 evaluations were identified. Thirty-nine states provided information about the individual studies that were completed. The details about the evaluation included: the genesis of the study, who performed the evaluation, the research techniques used in the study and the subjects chosen for study. In this section of the report, the unit of analysis is the individual study. Since some of the features were difficult to identify for some studies, the number of cases varies from table to table.

Based on state responses, less than half the evaluations were initiated by the state education departments; requirements imposed by external parties drove the state education departments in a slim majority of the cases. The findings show that more than

half the evaluations are initiated as part of oversight by legislators or federal programs operating through the state departments of education. The majority of evaluation studies were "farmed out" to consultants and academic researchers. Despite reliance on outside expertise, the majority of program evaluations were directed more to descriptive and process analyses and not to impacts or outcomes. The studies were largely founded on research methods that are highly unstructured and unscientific. The states are left with results that for the most part, carry no confidence that they are reliable. Even when study of program impacts was intended, the majority of studies used evaluation tools that could not ascertain the program effects.

Co-Sponsorship

Since one of the comments often made by respondents to explain the scarcity of evaluations is that there are no resources available to pursue them, respondents were asked if their state had co-sponsors for the studies they named. In the case of 51 studies, or about 14 percent of the evaluations, co-sponsors were used. The most frequent co-sponsors were other state offices, followed by the use of in-kind support from local universities in the form of faculty or graduate students. The use of co-sponsorship increased in the latter part of the period of study; 30 of the 51 were arranged in 2000 and 2001.

Choice of Evaluators

Each respondent was asked to classify if the evaluator responsible for each study was from the department, an outside consultant or an outside academic researcher. The research team learned that some states that designate some of their full-time employees as contract consultants to circumvent the hiring limits of the department. Persons of that description were considered departmental staff for this analysis. For this purpose, the term *consultant* is used to mean an independent unsupervised professional selected to perform an evaluation on behalf of the department.

Table 2 Choice of Evaluator

Type of Evaluator	Number of Studies	Percent*
In-House Evaluator	123	33.6
Consultant	160	43.7
Academic Researcher	113	30.9
Total number of studies	366	

* Percentages based on 366 identified evaluations. Figures do not sum to 366 studies because some studies used more than one type of evaluator.

The percentage of In-House Evaluators breaks down further to those studies done exclusively by department staff (27 percent) and those that combined internal and external evaluators (7 percent). Thus, about one third of all studies had internal involvement, but only about one quarter were the sole work of the department.

That two thirds of the studies were independent of the department is difficult to interpret with the data available. One might view reliance on outside expertise to reflect a concern

for impartiality in the evaluator. Perhaps it merely signals a lack of internal expertise. To gain further insight, we compared the predominate evaluator choices for each state.

States with only a few studies selected a single type of evaluator for most of their evaluations, though the choice varied by state. It would make sense to find an expedient arrangement and return to it as needed if a state were not planning to either expand its program evaluation activity or increase its own capacity to perform the work required. As the numbers of evaluations in a state increased, the pattern became more consistent; states either developed their internal capacity or diversified the stable of outside experts they could draw upon. In contrast, states that had completed numerous evaluations were more likely to use all three sources of evaluators. Whether this was in response to staffing limitations or reflected a more mature matching of needs to talents or both is unclear.

Project Resources

The studies varied widely in their scope, as reflected in the budgets allocated to them. Across the 207 studies for which information was provided, the figures ranged from a minimum of $2,000 to a maximum of $8,250,000. The median budget across all the studies was $189,355. Due to the large numbers of missing values, further analysis associating project budgets with the states' volume of evaluation activity, geography, or size of student population in the state was not pursued.

Reasons for Undertaking Evaluations

A key interest in this project was to learn more about the origins of the program evaluations that states undertook. For each reported evaluation, the respondent was asked to explain the motivation for the evaluation. All the reasons mentioned were recorded, though in only 20 studies were multiple responses offered. Reasons were offered for nearly 90 percent of the identified program evaluations (n=321). The responses were tabulated and are presented in the table below.

Table 3 Party Seeking Evaluation

Motivation for Evaluation	Number	Percent*
Required by State Legislature	101	31.5
Required by Federal Programs	75	23.4
Required by State Education Department		
by State Education Department Executives	22	6.9
by Division Staff	60	18.7
origin not specified within department	59	18.4
Other	4	1.2
Total number of evaluations with known origins	321	

* Percentages do not sum to 100% since multiple reasons were permissible.

As the figures show, about of the third of the studies were prepared in response to the state legislature. Most frequently, the evaluation was required by the legislature via the authorization process. Another quarter of the studies was required as a condition of receiving federal program funds.

These data also provide some insight into how program evaluations arise when departments of education initiate the work. Do they stem from an internal exercise of accountability (which is externally motivated from a close vantage) or do they reflect an interest by program staff in developing empirical feedback for self-managed program improvement? The choices mirror one of the key tensions in the perception of program evaluation: is it a tool for accountability or a tool of continuous improvement? Ideally, it can be both. Among the evaluations where the detail was available, staffs in the various program units were more likely to instigate evaluations than the executives of their agencies. It would be unwise to press the inference too far given the large numbers of department-generated evaluations whose genesis could not be more finely described. Regardless, less than half the studies were internally motivated, which suggests that

external mandates continue to be important for the generation of evidence on program performance.

Time Trends in Doing Evaluations

Respondents were asked the release dates for each of the studies they identified. Release dates are an objective measure of demand trends, despite the fact they are imperfect. They are imperfect for two reasons. First release dates are a proxy for start dates – the truest measure of demand – but studies may require differing periods of time to complete, thereby introducing some measurement error. Second, older studies are more likely to be under-reported due to imperfect recall, potentially leading to a spurious appearance of increases in more recent periods.

Table 4 Release Date of Evaluations

Year of Release	Number	Percent
Before 1997	14	4.8
1997	3	1.0
1998	14	4.8
1999	41	13.9
2000	38	12.9
2001	117	39.8
2002 and beyond	67	22.8
Total number of evaluations with known release dates	294	

For many of the evaluations included in this analysis, several reports are issued during the course of the evaluation. Multiple releases were reported for 16 evaluations. Another 14 evaluations were completed before the five-year period 1997-2001. These 30 studies were retained in other portions of the study but were not considered here. Another 67 program evaluations were underway during the five years of inquiry but not scheduled for completion until 2002 or beyond. Those that were completed between 1997 and 2001 show a strong increasing trend in the use of program evaluations in recent years compared to earlier. The dramatic surge in evaluations released in 2001 compared to 2000 cannot reasonably be attributed to imperfect recall of earlier years, and thus suggests an encouraging trend.

Evaluation Designs

Of the many research questions motivating this project, the most critical ones concerned the research methods and designs employed in the evaluations that states identified. Over recent years, the importance of design rigor and dependability of results has been elevated as policy makers scrutinize the existing education research base[7]. While several researchers have made noteworthy reviews of specific topics or groups of studies in meta-analyses, this is the first opportunity to examine the existing body of work produced by the states. In so doing, it is possible to identify the current capacity of the states to meet the challenge of higher expectations for quality and rigor in education research.

In 273 of the 366 evaluations identified by the states, information was reported on the type of evaluation and the research methods used to complete it. Table 5 presents the distribution of responses.

Table 5 Responses about Evaluation Type or Research Design

Response	Number of Evaluations	Percent of Total
Information Provided	273	75
Respondent Did Not Know	9	2
No answer	84	23
Total Number of Evaluations	366	100

[7] Ellen Condliffe Lagemann "What's the Matter with Education Research: Views from History". Askwith Education Forum, Harvard University Graduate School of Education, October 17, 2002.

Richard J. Shavelson and Lisa Towne, Eds. Scientific Research in Education, Washington, DC: National Research Council, 2002.

The loss of 84 studies without responses would be less troublesome were it not for the fact that a large number come from two states with the largest numbers of completed program evaluations. (Respondents from both states in question provided only the name and minimal details about each evaluation.) The unfortunate result is that we lack details from what could arguably be the best examples of state evaluation capacity. This in no way dims the picture that is painted for the rest of the states, however.

It should be noted that multiple approaches could be combined in a single evaluation. For example one state conducted a process and impact evaluation that included focus groups, customer satisfaction surveys and case studies as research methods. Accordingly, the percentages in the following tables should be considered independently.

Type of Evaluation Study We examined the characteristics of each of the 273 evaluations with known research designs. Where known, each evaluation was classified as cross-sectional, longitudinal or both. Details of the study design (whether it was a formative, process, or impact evaluation) were also recorded. Finally, respondents described the research methods used in each study to the best of their ability.

Only 134 evaluations were classified by their temporal design. Of these, 62 or 46.3 percent were *cross-sectional* in nature. In cross-sectional designs, evaluators examine subjects (e.g., students or schools) at one moment in time. For example, the drop-out rates in schools that implemented a retention program might be compared to those of schools that did not adopt the program (holding many other factors constant). Another 55

studies (41 percent) were *longitudinal*. Longitudinal studies incorporate a tracking component, examining the effects of programs over time. So-called pre-post evaluations are one sort of longitudinal study in which the outcome of interest is measured at two points in time (typically before and after a program is adopted or a student enrolls in a program), and the difference in measurement is compared to the change in non-participants. Seventeen evaluations (12.7 percent) were described as both cross-sectional and longitudinal, but it was not possible to ascertain if these evaluations were *repeated cross-sectional*—taking repeated cross-sectional snapshots in which the subjects could be different—or *panel designs*, in which the same individuals were followed over repeated periods.

The interview protocol asked respondents to identify the study designs of the evaluations they reported. A substantial majority could not distinguish between the various designs or their uses, or they had no idea what the terms were. Additional clarification, consultation with others in the department, and correspondence with consultants was needed in most cases. Even then, it was apparent that the terms are not defined consistently, even by professional researchers.

We classified evaluations as *formative* if the research focus was on validating a conceptual model or was focused purely on the operational improvement of a program. Such studies are common in the early years of a program. *Process evaluations* also look at the mechanics of programs by studying several sites, relating inputs and operations to outcomes in order to identify minimum operating requirements or critical success factors.

A study was coded as *Impact Evaluation* if its focus was to discern the effectiveness of a program in achieving its intended outcomes.

Many of the evaluations combined research designs in a single study. Accordingly, many were recorded in multiple categories. The resulting proportions indicate the share of all the evaluations that used each design, but the proportions are not summative. The figures appear in Table 6 below.

Table 6 Evaluation Designs

Evaluation Type	Number of Evaluations	Percent of Evaluations with Known Designs*
No answer	84	n/a
Don't Know	9	n/a
Formative Evaluation	102	37.4
Process Evaluation	136	49.8
Impact Evaluation	155	56.8
Total with Known Designs	273	
Total Evaluations	366	

* Percentages are not additive because a single evaluation could employ multiple evaluation types.

Restricting the calculations to the subset of program evaluations where the designs were known, slightly more than half the studies sought to examine the impact of the programs under review. Nearly half also were structured as process evaluations; they sought to learn more about the way program sites were operating either in comparison to each other or against some program model. Slightly more than a third of the studies were formative in nature; they were structured to look at new or young programs and provide information to refine the program elements or implementation strategies.

Using the grouping of states by their activity level presented earlier, the tabulations were further examined to see if having more experience with program evaluations changed the types of evaluations that were pursued. The numbers are reported as the percentage of studies done in states in each group for which the evaluation type was known.

Table 7 Evaluation Type by Extent of Evaluation Activity
(Values are proportion of states included in each row.)

State Evalution Activity	Formative Evaluations	Process Evaluations	Impact Evaluations
1 to 5 evaluations	32.4	58.9	37.8
6 to 10 evaluations	31.3	40.5	58.8
11 to 20 evaluations	33.3	64.7	58.8
21 or more	59.3	53.7	62.9

Note: Percentages are not additive because a single evaluation could employ multiple evaluation types.

The proportions show that in states with little evaluation experience, a smaller share of the studies are impact evaluations. Another point worth noting, is that the states performing a lot of studies appear to take a more comprehensive approach to their evaluations, combining evaluation types more than in other groupings.

Research Methods The responses from the states on the research methods used for their evaluations was perhaps the most revealing of the entire project. Again, the base is the 273 program evaluations for which research design parameters were reported. The aim of this portion of the study was to see what approaches states used to design or collect data. Each of the approaches offers a different degree of objectivity and control of potential sources of bias. While the list mixes design and data collection methods, each choice has come to be considered a separate technique for conducting evaluations. They are used here in that larger context. The tabulations appear in the table below.

Table 8 Program Evaluation Research Methods

Evaluation Methods	Number of Evaluations	Percent of Evaluations with Known Designs*
Random Assignment	19	6.9
Quasi-Experimental Design	40	14.7
Trend Analysis	69	25.3
Focus Groups	73	26.7
Satisfaction Surveys	118	43.2
Total with Known Designs	273	

* Percentages are not additive because a single evaluation could employ multiple research methods.

The research methods are listed in decreasing order of research rigor, based on the work recently released by the What Works Clearinghouse / Campbell Collaboration[8]. A number of insights emerge from the tabulation. First, the proportion of all identified evaluations studies that employ a comparison group of any kind for the purposes of attributing program effects is very slim. Of the available research methods, only random assignment and quasi-experimental designs control for program effects – they identify individuals who are not participants of the program in question and compare their outcomes to those in the program. They differ in the degree to which they control other potential influences: random assignment produces the purest draw of cases and controls. These studies are the only ones that can reasonably defend their findings as reliable. Less than 22 percent of the evaluations met this criterion.

[8] What Works Clearinghouse, *Study Design and Implementation Assessment Device*, (Study DIAD), Version 0.6, and *Cumulative Research Evidence Assessment Device* (CREAD), Version 0.6. (Rockville: What Works Clearinghouse, 2003).

Second, a large number of studies rely on the self-reported status of the program participant, parent, or teacher to gather information about the program in question. While focus groups and satisfaction surveys methods have some utility in offering perspective from different constituencies, these sources are questionable as objective and unbiased reporters.

Third, the value of trend analysis depends entirely on the comparability over time of the data being compared. Simple comparisons across time may inadvertently capture changes in outcomes due to shifts in factors other than the program of interest. Since many studies rely on local samples contrasted with reference statistics drawn from external sources, this technique is especially vulnerable to faulty interpretation. The inference from these figures is that states are basing their judgments about program performance on shaky ground.

In an effort to better understand the data, three additional analyses were performed. These data were first examined against the overall level of activity in each education department to see if differences exist. The general idea was to test if departments with greater volumes of studies used more rigorous or sophisticated research methods. The results were surprising – there were no notable differences. Perhaps if the two states with known large numbers of studies had been incorporated, the results would have been different, but based on the available data, volume of activity does not correspond to clustering in the more rigorous research methods.

An alternate analysis considered the possibility of learning over time. Since the concentrated attention to the quality of education research has occurred only over the past couple of years, perhaps the states chose research methods differently in the last years of the study period. Shifts in the proportions of studies using each of the methods would be apparent if states decided to make their program evaluations more rigorous. The choice of methods was examined by year; in this analysis we included the studies with release dates of 2002 and beyond to provide the most current picture possible. The results are displayed below in Table 9.

Table 9 Research Methods Over Time

(Figures are percentages of all studies completed in the year.)*

Row Percentages Using Each Method

Release Date	Number of Evaluations with Known Designs	Random Assignment	Quasi-Experimental-Design	Trend Analysis	Focus Groups	Customer Satisfaction Surveys
1977 and earlier	16	12.5	12.5	18.8	18.8	0.3
1998	11	0.0	18.2	18.2	27.3	63.6
1999	20	0.1	0.3	0.2	0.3	0.2
2000	27	0.1	0.2	0.4	0.4	0.4
2001	104	6.7	12.5	26.0	28.8	56.7
2002 and beyond	55	7.3	12.7	23.6	29.0	41.8
Date unknown	40	5.0	10.0	0.3	15.0	0.3

* Percentages are not additive because a single evaluation could employ multiple methods.

The table clearly demonstrates the frailty of recall as time passes—details from the earlier years were scarcer than from recent times. However, if the evaluations with known research methods are typical the evidence does not suggest a shift to more rigorous methods.

The preceding tables on evaluation type and research methods beg the question of how various evaluations sought to demonstrate the effectiveness of the programs they examined. The question is especially germane in light of the increased concentration on school and student performance and the creation of accountability systems in almost all the states[9]. With consistent measures of student and school performance in place in every state, the next logical step would be to focus on identifying which programs contribute positively to good achievement and which ones do not. The large number of impact

[9] Eric Hanushek and Margaret E. Raymond (June 2001), "The Confusing World of Educational Accountability", *National Tax Journal*, volume LIV, no. 2.

studies suggests these questions are being asked. Given the large numbers of impact evaluations and the relatively small number of random assignment and quasi-experimental design evaluations, how did the states attempt to estimate the effect, impact or outcomes of their programs? The results appear in the table below.

Table 10 Impact Evaluations by Research Method

Evaluation Methods	Number of Evaluations	Percent of Impact Evaluations
Random Assignment	10	2.6
Quasi-Experimental Design	33	6.4
Trend Analysis	49	31.6
Focus Group	42	27.1
Satisfaction Surveys	59	38.1
Other Approaches	4	2.6
Total Number of Impact Evaluations	155	

The discontinuity between evaluation intent – to estimate the effect of the program in question – and the methods used to gauge those effects is remarkable. On the one hand, significant resources are involved in the funding of education programs, so the interest in impact evaluation is both appropriate and reassuring. On the other hand, the majority of evaluations used to estimate the effect of those programs have significant shortcomings. The resources devoted to these evaluations could have been better spent on more reliable studies. Worse is the real possibility that they have engendered a false sense of reality. Since we do not know the ultimate influence these evaluation results may have had, we can only hope that future evaluations will be planned with greater attention to maximizing confidence in the findings.

Program Areas of Evaluation

For each program evaluation that was identified, state officials were asked to specify the topic of the evaluation. All the responses were recorded verbatim and subsequently clustered into related groups. The resulting frequencies appear in the table below.

Table 11 Program Evaluation Topics

Topic Group	Number	Percent
Education Programs for Special Populations	68	18.8
General Education Programs	67	18.6
Non-Core Curriculum & Support Services	61	16.9
Other	40	11.0
School Reform Programs	39	10.8
Teacher/Administration/Parental Programs	31	8.6
School Achievement	21	5.8
Early Childhood	17	4.7
Student Achievement	15	4.2
Operations/Facilities	2	0.6
Total	361	

The distribution shows that a wide array of programs has been studied in the past five years by the states. Perhaps most surprising is the relatively small number of studies that relate to teacher preparation and early childhood programs. Both have received considerable program support and public attention, but have not seen corresponding study of their performance.

The largest categories have been general education (which includes any program that is available to all students), education programs for special populations, non-core curriculum and support services such as health education or violence prevention, and school reform programs. Given the strong policy attention in these areas in recent years, the concentration of evaluation effort in these areas is appropriate and timely.

However, different policy makers may have different areas of focus. State legislatures may want evaluations of programs that affect every student, where the federal actors may prefer to learn about the special populations that are the focus of federal programs and funding. To investigate if differences exist, we cross-tabulated the evaluation topic for each study with the source of the evaluation's initiative. The results appear in the following table.

Table 12 Evaluation Topics by Party Initiating the Evaluation

Topic Group

Party Initiating Evaluation	General Education Programs	Education Programs for Populations of Interest	Non-Core Curriculum & Support Services	Student Achievement	Teacher/ Administration School Achievement	Teacher/ Administration /Parental Programs	Operations /Facilities	Early Childhood	School Reform Programs	Don't Know	Total
State Legislature	28 27.72%	22 21.78%	16 15.84%	4 3.96%	3 2.97%	3 2.97%	2 1.98%	3 2.97%	20 19.80%	0 0.00%	101 100.00%
Federal Programs	37 49.33%	7 9.33%	13 17.33%	0 0.00%	0 0.00%	2 2.67%	0 0.00%	6 8.00%	10 13.33%	0 0.00%	75 100.00%
State Education Department Executive	4 18.18%	2 9.09%	4 18.18%	3 13.64%	6 27.27%	0 0.00%	1 4.55%	0 0.00%	2 9.09%	0 0.00%	22 100.00%
State Education Department Division Staff	18 30.00%	8 13.33%	13 21.67%	3 5.00%	4 6.67%	6 10.00%	2 3.33%	4 6.67%	0 0.00%	2 3.33%	60 100.00%
Origin not specified within Department	22 37.29%	8 13.56%	11 18.64%	2 3.39%	3 5.08%	6 10.17%	1 1.69%	3 5.08%	1 1.69%	2 3.39%	59 100.00%
Other	3 75.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	1 25.00%	0 0.00%	4 100.00%
Don't Know	0 0.00%	0 0.00%	1 33.33%	1 33.33%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	1 33.33%	0.00%	3 100.00%
Total	112 34.57%	47 14.51%	58 17.90%	13 4.01%	16 4.94%	17 5.25%	6 1.85%	16 4.94%	35 10.80%	4 1.23%	324 100.00%

The results illustrate that indeed different actors have different foci in their evaluation interests. State legislators have general education and education programs for special populations as their top interests. The spate of legal challenges to states' school desegregation policies may explain some of the interest. The areas of greatest evaluation focus initiated by federal authorities are general education programs (which include reading and mathematics instruction), school reform efforts and the programs that arise out of the Department of Education's special initiatives. Executives in state education departments focus most on school achievement studies, most relating to the evaluation of state accountability systems or remedial intervention with poorly performing schools. Since the priority areas do differ, it will be important to have state legislatures and the federal department seek evaluations in the areas that do not coincide with state priorities if there is to be an accumulation of evidence about a wide range of programs.

V. Departmental Management of the Evaluation Practice

In addition to designing and executing program evaluations, it is necessary for states to generate the studies, provide oversight and review of the deliverables, vet the results, disseminate the results, and assimilate the findings in future program decisions. We refer to these associated tasks as evaluation management. In many ways these aspects of education department activity are as important as the individual studies themselves, since they can potentially affect the number, quality and ultimate utility of the investment in evaluations. The survey of state education departments collected information on states' evaluation management. The findings presented in this section show that even in states with dedicated research and evaluation staff, few states follow regular procedures. Often states patch together ad-hoc arrangements, with little attention to on-going quality assurance activities. Dissemination of results appears to occur at the discretion of the departments. There are no formal means to consider the findings in future decisions.

Organizational Emphasis

One obvious gauge of the importance placed on measuring program performance by states is whether the education department is structured to include a division with that as a prime mission. Organization theory suggests that the allocation of scarce resources reflect the unobserved priorities and values of the executives of each organization[10]. A

[10] Richard H. Cyert and James G. March, Behavioral Theory of the Firm. (1963) Englewood Cliffs, NJ: Prentice Hall, p. 272.

decision to make research and evaluation an official function of the department gives the operation official sanction and standing. Official designation also helps the division managers' ability to vie for additional resources to increase the visibility and influence of the division, a natural drive within any bureaucracy.

States were asked if their education department had a separate internal group that focused on research and evaluation activities. Since nine of the states do no evaluations, it is reasonable to assume that they have no official evaluation function within their organizations. Eighteen states responded affirmatively, amounting to 36 percent of all states and 44 percent of the states with evaluation histories. Twenty-three states do not have separate divisions within their departments.

The fact that the evaluating states fall roughly into halves on the matter of officially designated function within the larger department prompts the question of the impact of such status on the output of the states. As a first cut at the question, a t-test of the association of having a separate research and evaluation division and the number of evaluations conducted by each state showed no significant difference.

Staffing

A key motivation for this project was an interest in the availability of staffing and expertise. A variety of skills are needed to plan, organize, oversee and review program evaluations in state education departments. The variance in the numbers evaluation of studies across states might be explained by differences either in staff or in the necessary research and evaluation expertise across departments. Despite observed differences in the human resources and commitments of time to evaluation activities, which are explained below, our analysis found no significant association between staffing with output.

Recall that nine states have no recent history of evaluations. This leaves 41 states for which staffing for evaluations are relevant. Of the 18 states with research and evaluation divisions, 17 provided staffing information; seven of the 17 also draw on personnel in other divisions to assist with evaluations. The remaining 23 states do not have research and evaluation units. Several states explained that in recent years, staff with research backgrounds had been redeployed to support initiatives in testing, school accountability, or data systems. No information on staffing was obtained from five of these states. We were able to obtain detailed information about staffing from eighteen of these states: 14 states use staff in other divisions on an ad-hoc basis, and 4 states use no internal staff at all.

States were asked to identify the individuals in the department who participate in its evaluation activities. From the responses, it was possible to tabulate the number of positions in each state. These results are summarized in Table 13.

Table 13 Personnel Working on Program Evaluations in State Education Departments

		Number of Positions		
		---------- Range ----------		Average
	Number of States	Minimum	Maximum	
SED Has No Research and Evaluation Unit	23			
▪ No SED Staff Involved (4 states)				
▪ Personnel in Other Divisions (14 states)		0	12	4.2
Total Staff Positions		0	12	4.2
SED Has a Research and Evaluation Unit	18			
▪ Only Uses Personnel in R and E (10 states)		2	13	4.4
▪ R and E Staff and Personnel in Other Divisions (7 states)		1	12	6.2
▪				
Total Staff Positions		2	18	7.4

Based on the number of personnel involved in evaluation activity, it would appear that SED's with a separate Research and Evaluation unit have on average a larger pool of talent. 7.4 compared to 4.2 positions. The boost in available manpower is substantial, especially in departments that draw from both research and evaluation units and other divisions to work on evaluation matters.

One might suppose that the difference in personnel resources might create a larger wealth of experience, but there was so significant difference in the average years of experience among states that had larger resources than smaller ones.

On a related point, there was virtually no variation across states in the reported prevalence of a number of research and evaluation skills. However, the skill categories (research methods, survey methods, program evaluation, qualitative methods and applied statistics) were broadly construed, leaving specific interpretation to the respondent, so it is possible that the data supporting the finding is inconsistent.

Secondarily, we found no significant association between the total number of individuals and the volume of evaluations completed over the five year period of investigation, although the coefficient of $r=.31$ approached significance at $p<.06$.

What we found instead was a surprise: the difference in numbers of positions narrows if one looks at the total level of effort devoted to evaluation activity. Each state was asked to record the percentage of time each individual actually spent on evaluation in the course of a year. These figures were used to calculate the total full time equivalents (FTE's) devoted to program evaluations for each state. The results were then averaged within the groups laid out in the previous table. The results appear in Table 14.

Table 14 Personnel Resources Working on Program Evaluation in State Education Departments in Full Time Equivalents (FTE's)

		FTE's Range		
	Number of States	**Minimum**	**Maximum**	**Average**
SED Has No Research and Evaluation Unit	23			
▪ No SED Staff Involved (4 states)				
▪ FTE's in Other Divisions (14 states)		.05	4.5	.93
Total Full Time Equivalents		.05	4.5	.93
SED Has a Research and Evaluation Unit	18			
▪ FTE's in Research and Evaluation Unit (17 states)		.32	6.4	1.7
▪ FTE's in Other Divisions (7 states)		.05	2.5	1.2
Total Full Time Equivalents		.32	6.4	2.3

The figures in Table 14 are noteworthy for several reasons. Overall, not much investment is made by states to pursue program evaluations. While a very few states devote substantial personnel resources to these matters, the averages show that typically states do not provide much staffing. Even within formal units devoted to research and evaluation, the allocations are routinely small, less than 3 full time positions. It is hard to envision a robust program evaluation practice for a state being supported by such limited investments of staff.

This finding is consistent with anecdotal evidence provided by several of the respondents. Largely due to the skill set among research and evaluation staff, their recent activities have been directed to the design and adoption of state accountability systems for schools and students. If the survey is repeated in the future, the allocations of time to program

evaluation may increase as states transition from implementation to maintenance of these accountability systems.

Second, it is not surprising that on average states with research and evaluation units put more resources to bear than those states without. The difference is 1.39 FTE's on average, a finding that is statistically significant at the $p > .01$ level ($t = -2.8375$). What is perhaps the more surprising finding is that the difference in FTE's does not correlate with a difference in output: there is no significant difference in the numbers of evaluations across the range of total FTE's.

Use of Outside Expertise

One plausible explanation for the low staffing ratios is that the states elect to have their evaluation work done by outside consultants or academic researchers. If this were the case, one might expect a negative correlation between the allocation of internal manpower and the propensity of a state to use outside expertise on the evaluations they perform. The proportional use of outside manpower was measured as the proportion of evaluation studies in each state that used outside consultants or academic researchers. Of the 41 states that did evaluations, information on staffing for particular studies was available from 39. Only one state performed all their studies in-house. The remaining states ranged from using outside experts 28 percent of the time to 100 percent of the time, with an average of 76 percent across all the states. The relationship between the use of

outside expertise and internal manpower was only weakly negative ($r = -.09$) and not significant. Thus, the low levels of internal staffing do not appear to be systematically offset through augmentation.

Funding for Evaluation

Obtaining good information about the budgets for past evaluations turned out to be one of the thornier sections of the survey. Many of the survey respondents exerted great effort to provide the information requested, but in this area, their efforts were not as fruitful as in others. Over half of the 39 states that provided details about their evaluations found it difficult to retrieve budget information for past studies or declined to provide the information. In six states, no budget figures were provided for any of their studies; for another 20 states, the respondents supplied as much as they could but the picture is incomplete. Only 8 states, all with small volumes of activity, gave complete details. So, we were able to obtain budget information for varying proportions of their evaluation studies from 28 states.

Lack of recall was compounded by legitimate differences in the ways states budget for the evaluations they perform. For example, in many states every evaluation is separately approved and so all the budgetary resources are identified explicitly. In others, resources for at least some of the evaluations are included in the basic departmental appropriation for the department or are secured from the discretionary budget for the department. This

would make sense for states with research and evaluation units, and the data indicate that 8 of the 18 reported that some of their evaluations were not budgeted for separately. Six of the states without research and evaluation units also reported embedded budgeting for some of their projects.

Despite the limitations of the data, budget information was provided for 200 studies from 28 states. The budgets for individual studies were totaled, yielding a basis for comparing the states on their ability to secure funding for studies over several years. The smallest total was $2,000 while the maximum was $8,250,000, with a median across the 28 states of $189,355.

It would be unwise to push these data too far, but one interesting finding emerged from the analysis. On average, the states with some degree of embedded budgeting secured over $1,750,00 for evaluation (over and above the undifferentiated department funding) compared to the average of $1,113,000 for states that budgeted on a project by project basis. The difference was about 55 percent higher, but the difference is not statistically significant, due to the wide variation around the averages. The specific causal relationships are not apparent, but one possible explanation is that routine use of department resources for evaluation creates a foundation of familiarity and confidence that facilitates acceptance of larger projects as they arise.

Administrative Procedures for Evaluations

The ways that states manage their evaluation activity can be as important as the quality of the studies themselves. To illuminate these important mechanisms, state respondents were asked to explain the procedures for soliciting and reviewing proposals, reviewing work products and final reports, and disseminating the results of the completed evaluations. We received information from 36 of the states. The information they provided shows that these management components of the evaluation practice are handled for the most part by ad-hoc procedures. Further, there are no procedures for capturing the experience gained from administering prior evaluation studies to improve the quality and appropriateness of future ones. The combination results in a degree of "reinventing the wheel" that hampers the opportunity to incorporate evaluation of program performance as a regular part of operations.

Requests for Proposals At face value, it might appear that the Request for Proposal (RFP) step in the evaluation process would be fairly straightforward. But in order for an RFP to generate high quality proposals it must do all of the following: correctly identify the question(s) for study, provide sufficient information to identify the breadth and variety of the program under study, articulate the limitations that evaluators should anticipate and accommodate in their proposals, explicitly describe any "non-negotiable" elements of the project, describe available resources for the project, and outline the desired format of the proposal submissions. Optional features also include: how the

various components of a proposal will be weighed in a final decision, description of earlier evaluations of the program, or suggested approaches or designs.

In order to prepare an RFP with the necessary information, it is necessary for states to have done enough preliminary work on the evaluation to be able to articulate each of the requirements. Further, there needs to be a complement of program expertise and sufficient technical skill to be able to ascertain whether a desired evaluation is even practicable given enrollments, program size, geographic dispersion, feasibility of control groups and so on. More detailed issues such as plausible research designs, sampling strategies, feasibility of surveying, or various methods of analysis also need to be thought through prior to preparing the RFP.

These same skill sets are needed to review the submissions to assess the match between proposed evaluations and the objectives of the project.

The analysis shows that all 36 states that provided responses use internal staff to prepare their RFP's. In 31 states, the process is handled largely by staff of the program to be evaluated. In eight states some additional technical support is obtained from senior members of the department, the contracting office or the legal department. Only in four states were outside experts consulted for any of the RFP's. Surprisingly, among the 18 states with research and evaluation units, five of them leave the RFP development to program staff entirely. So while the programmatic aspects of RFP's appears well covered, the more scientific or technical matters are handled only to the extent that the

necessary expertise is extant in the program unit. This finding goes a long way in explaining the dearth in rigorous study designs among the evaluations that states identified.

Several states identified better procedures for reviewing proposals. The choices of internal versus external reviewers each have benefits and limitations. Internal reviewers may be familiar with the program being evaluated or have a sense of the organizational contexts surrounding the evaluation. Their potential drawback is a lack of objectivity. The converse applies to external reviewers. (As a reference point, the U.S. Department of Education uses mostly external reviewers.) Thirty-four states provided information on their proposal review practice. Thirteen states (38 percent) reported that they have specific approaches to review proposals; of the 13, nine states use a common scoring rubric and four have formal review policies and procedures to assure impartial and consistent consideration of the submissions. In 21 states (62 percent), the process is informal, ad-hoc or "self-guided".

Review of work products and deliverables Another vital way to assure the appropriateness and quality of the evaluations that are completed is through the review of work products and deliverables. Thirty four states reported on their procedures. In this management area the general practice reported by states is highly fluid, without many mechanisms to assure that studies are well executed and of sufficient quality to use in future decisions.

While it is clear from the previous section that the quality of the evaluations done in the states varies widely (with corresponding variation in the confidence their results can engender), it appears that an appreciation for these distinctions is not well understood. No respondents reported that their review process examined the overall reliability of the studies they authorized. If it is mostly the case that education policy makers accept all studies to be of equal quality, then a considerable disservice is being perpetuated, not only to the evaluators whose work is in fact superior, but also to the public on whose behalf decisions are made based on questionable findings.

Sixteen states indicated that the process is non-specific and varies by project. Several states review the analysis or the final results using internal or external staff. Others compare final products with contracted deliverable requirements. Discussions and review with the evaluators or an open question period for anyone who reads the evaluation are also used. The viability of these alternatives clearly depends on the substantive knowledge – both programmatic and technical – of those doing the review. In two states, every evaluation report is presented to the department executives. Only one state uses an expert review process, asking individuals from outside the department to review the final reports for completeness and accuracy.

Of greatest concern were the findings from three states. In two states, staff who administer the programs are given the exclusive responsibility to review the final evaluation reports. In the third, the staff negotiates with the evaluators on studies that are politically sensitive or highly visible to make adjustments. In all three cases, there are

significant questions of moral hazard. The integrity of the evaluation could be at risk if interested parties are left to judge the merits of the work, especially in cases of unfavorable results.

Dissemination of Results One of the primary reasons to perform evaluations is to create evidence about program impacts to inform future decisions. Another motivation is to exert a degree of program accountability to maximize the effectiveness of investments during the current funding period. In both scenarios, the usefulness of evaluations is determined in large measure by how widely the results are shared.

Clearly, the departments of education themselves have many good uses for the evidence that evaluations produce: program improvements and the fine-tuning of implementation strategies are examples. Even within departments of education, a degree of accountability can be exercised over programs through executive review of evaluation results. This oversight cannot occur, however, when results are generated and shared only within the group that manages the program being evaluated.

Likewise, since most states craft policies multilaterally with input from legislators, governors, interest groups and the departments of education, each of those constituencies has an interest in seeing the evidence about program performance. To be sure, unfavorable results can escalate political problems for a program, but such vetting is both appropriate and required by the pluralistic political structures of state government.

Where results are not shared, the risk is elevated that personal or parochial interests will be placed above the public interest.

There is a third use to be realized from sharing the results of evaluations – doing so supports the development of knowledge both about programs and about evaluation practice outside the state in which the study is done. The accumulation of evidence about the impacts of programs is behind recent federal initiatives to gather, analyze and disseminate findings from education evaluations for the benefit of all decision makers.

Against this backdrop, 38 states described their actual dissemination practices. Their responses are displayed in the table below.

Table 15 Dissemination Practices

Method	**Number of States**	**Percent**
Mailings	36	94.7
Formal list	21	55.3
Ad-Hoc list	15	39.5
Post Results on Web Site	16	42.1

Thirty-six rely on mailings to distribute evaluation findings. Of these, two states volunteered that they will release reports only when required to do so by legislative or federal requirement. In the remaining states the atmosphere appears more cooperative, though no state reported that they routinely distribute all their reports. Twenty-one states or 55 percent of those responding use a formal mailing list: 20 include the legislature and

16 include the governor's office. Fourteen states include county/regional or local education agencies in the release of studies.

The mechanics of distributing evaluation reports might constitute an impediment to fulfilling the larger functions of dissemination. With the virtual ubiquity of Internet access and the universal hosting of Web sites by education departments, one low cost and convenient means of distribution lies in posting reports on the department's Web site. Only about a third of those reporting – 16 states or 42 percent – use their sites for this purpose. Given the dramatic economies that this option presents, it is surprising that more states do not avail themselves of it.

Incorporation of Results into Future Decisions The ultimate worth of an evaluation is the extent to which it influences future action. The ways in which state education departments use the results of their evaluations offer another view of the current role that empirical evidence plays in state education departments. Information was shared by 36 states on their procedures to include the results of evaluation studies in future decisions. The data reveal that the majority allows considerable latitude to consider or disregard evaluation findings. While 19 states (53 percent) report some sort of formal mechanism to factor results into later program decisions, in 11 of them the process is followed inconsistently. In the remaining eight states with formal procedures, senior executives review the results of each evaluation, factor the results into future reports to the legislature, or adjust their internal management plans based on the findings. Twelve

states (33 percent) use only informal procedures to build on evaluation findings, and five states (14 percent) use other means.

Research Synthesis

The findings presented in the previous sections each considered the associations of a variety of characteristics about the education departments with the numbers of evaluations they produced. The original research design for this project called for multiple regression analysis to incorporate the set of descriptors into a consolidated model of state evaluation capacity. The aim of the comprehensive analysis was to consider simultaneously the contributions of staffing, budgets and the like to the observed activity across states.

With the limited number of observations available, it is difficult to separate out the various influences on evaluation activities in the states. We have observed several suggestive correlations between evaluation outcomes and various characteristics of the state structure. Because the underlying characteristics of states tend to be correlated with each other, it would be useful to assess the separate influence of each.

Nonetheless, when we turn to the multivariate modeling, we cannot adequately distinguish among the factors. None of the factors is statistically significant from zero in multiple regressions that include the separate factors simultaneously. This lack of

significance could arise because none of the factors truly influence the outcomes. However, because of the limited data, it is more likely that we simply do not have enough available information to be confident of the separate influences.

Another point bears mention. The observed characteristics – staffing and budget and so on – may actually reflect more complex and as yet unmeasured explanations of what is occurring in states; in this sense they are the "reduced form" equations for more sophisticated structural models of state behavior. What we observe in the states may signify different underlying dimensions, such as perceptions of the role of evidence, estimates of the political costs, and benefits of producing explicit performance information, staff quality, availability of expert evaluators, or the organizational attitudes about primacy in the policy arena. These other relationships were beyond the scope of the current project but are worthy of further examination.

VI. Visions of Future Development

The final area of the survey asked respondents to identify their three top "wish list" choices for departmental changes in the area of program evaluation. Thirty-seven states offered over 100 suggestions. The responses focused generally on ways to enhance the quantity and influence of program evaluations. While the suggestions may seem self-interested coming from the departments' most knowledgeable authorities on evaluation, that risk is tempered by the fact that these individuals also are aware of the potential role that solid empirical evidence about programs can play in improving the effectiveness of their departments. The responses were analyzed and appear in the table below. Within each general category, we identify the range of specific suggestions.

Table 16 Evaluation Wish List

Suggestion	Number	Percentage of Suggestions	Number of States	Percentage of States
Resources	28	26.4	19	51.4
Institutional Support	22	20.7	15	40.5
Staff	15	14.2	14	37.8
Training/Professional Development	9	8.5	9	24.3
Evaluations for Specific Programs	15	14.2	8	21.6
Time	5	4.7	5	13.5
Better Political Environment	2	1.9	2	5.4
Basic Research	2	1.9	1	2.7
Strategic Plan for Evaluation	1	0.9	1	2.7
Trends/Statistics	1	0.9	1	2.7
Other	6	5.7	6	16.2
Total Number of Suggestions	106			

37 States provided responses.
Note: The figures are not additive because states could make more than one choice.

The two areas that prompted the most recommendations for the most states were the desire for more resources for evaluations (26 percent of suggestions from 19 states or 51 percent of those responding) and better institutional support for program evaluations in current operations (almost 21 percent of the suggestions coming from 15 states or 40 percent of respondents). Nearly as popular across the states was the request for more qualified staff: 14 states mentioned this need. Finally, nearly 15 percent of the suggestions were to develop evidence on the performance of specific programs. These suggestions came from 8 separate states.

One positive implication of these results is that there is already a degree of recognition in many states that improvements in capacity, quality and influence are needed. The fact that so many states made similar recommendations also suggests that there may be a common approach to their suggestions. Aside from the infusion of resources, it is conceivable that states could cooperate to bring greater expertise to bear on planned studies. They could also explore the feasibility of sharing a common design for evaluation of common programs. They might even share the expense of professional development for the staff members that already are engaged in evaluation work.

Summary of Findings

Great weight is put on states to craft policies and programs that will improve the education of America's public school students. There is no doubt that educators and other policy makers are fervent in their motivation to improve education outcomes for public school students and take aggressive action towards those ends. Policy makers and educators alike need regular and reliable information about how their programs are performing if they are to make wise adjustments to existing programs. But without the critical evidence and feedback about performance that solid program evaluation delivers, a vital navigational tool is lost. To capitalize on the investments states have made in accountability systems – which capture only aggregate outcomes – states and schools need to be able to discern what elements of their efforts are responsible for gains and which are not helpful. The majority of states currently lack that ability. The findings presented here demonstrate that program evaluations are infrequent, of dubious methodological rigor, largely not about outcomes, not reviewed for quality, and are not formally considered in the later decisions.

This project was in many ways a first effort, but the findings it produced are illuminating and important. The research shows in striking ways that state education departments are not doing an adequate job of requiring, executing or using evaluations of their programs as a critical component of policy making. About a third do little or no evaluation; about a third do one to two evaluations per year; and a third evaluate their programs more extensively. The survey with state education departments identified 366 evaluations conducted by them or under their auspices in recent years. It would not be surprising to

learn that other program evaluations were omitted from the compilation, given the large number of states with decentralized responsibility for assessing program performance. But, based on available data, it is clear that states do not regularly seek review of the programs they put in place.

In one way, the difficulty in capturing the universe of evaluations is itself an interesting finding: if after six months of effort we missed some number of studies, it is safe to assume that they are inaccessible to others as well. The evidence provided by program evaluations can only be useful if it is used in charting the future course of policy. Any buried studies are at best useless and at worst obstruct rational decision-making.

The findings show that the state education departments initiated fewer than half of the evaluations. Legislatures and the U.S. Department of Education filled in the remainder and will continue to be an important source of motivation for evidence on program performance. Yet, the hope remains that over the long run, state education departments will make performance feedback a regular part of department operations and oversight. The associated expense is similar to insurance: it weighs against the significant chance that program dollars are being spent ineffectively year after year.

Where evaluations were undertaken, we find an alarming disconnect between intention and action. The results show wide variance in the capabilities of state education departments to undertake reviews of the performance and effectiveness of their programs.

The inferior methods of many of the studies that we examined cannot offer any solid basis for decisions, contrary to the increasing call for "research based policies." Even where a state sponsored one of the few identified studies that employed a methodology that lends confidence to the results, it is not the typical practice for that state. If a review of program initiatives is trivial or of poor quality, then decisions about the continuation of programs must be founded on other, less objective, grounds.

It is disconcerting that lots of resources are spent on program evaluations that cannot discriminate true program effects from a host of other potential explanations. And even where the respondents we interviewed may be knowledgeable in the difference, the prevalence of inferior quality evaluations strongly suggests that such recognition is not widely shared in their home organizations. If state education departments are to be active parties in judging the worth of their programs (as a component of good agency management a well as internal accountability), there will need to be significant changes in the quality standards used these departments. The recently released draft standards by the What Works Clearinghouse offer a universal and objective set of standards that states can use.

What is the effect of ineffective evaluations? The existing body of evaluations make it extremely difficult to answer questions about program operations or impacts. It is not possible to tell from the information on hand if poor quality evaluations have affected specific programs. It is easy to imagine such a scenario, but it is also conceivable that the body of work has been accurately assessed as suspect by policy makers and therefore

discounted in the decision process. In either case, the purpose of examining program performance has not been realized.

The research results were less conclusive than expected about the underlying causes of the differences in states' capacities. Significant differences in staffing and budgeting were found to be associated with different levels of activity, but when combined with other factors in a more comprehensive estimate of differences in output, none of these effects were significant. A more comprehensive exploration of the dynamics that operate within state education departments may be needed before such insights are possible.

In almost every state included in this research effort, department respondents acknowledged the need for improvement in their organization's ability to assess program performance critically. They recognized the need for higher priority to be devoted to empirical evidence. The perception of need for more resources to launch more evaluations was widespread. States also need professional development to enable staff to do better work.

The lack of ability to explain what causes the observed differences between states in their level of effort is unfortunate but is perhaps less important than the normative question that the overall findings raise. What should states be doing going forward? Giving the benefit of the doubt, it could be that to this point state education officials were unaware of the deficiencies in their departments. Since this report will be distributed to each state education department as a courtesy for participating in the study, they will now have a

clearer picture of the state of affairs. The respondents from the state education department were themselves able to identify steps that could make dramatic strides in the capacity of states to review the effectiveness of the programs they sponsor. Whether states move to adopt such changes will reveal a lot about the underlying motivations and cultures in the departments. Their individual responses to these findings presents an excellent opportunity for further study.

Future Directions

This analysis shows a clear need for State education departments to continue development of their capacity to provide regular and rigorous assessment of their programs. States differ in their current levels of enterprise, and so will need differing degrees of needed development. Regardless of states' current abilities, however, it is clear that leadership will be needed from legislatures, chief state school officers, and the U.S. Department of Education if progress is to be made.

Perhaps the largest challenge is to help education agencies see that evaluations have little value if they occur in isolation. An increase in expertly designed and executed program evaluations is certainly desirable, but more is needed. A larger set of organizational elements within departments must also be aligned before and after an evaluation is performed. Prior activities include setting an evaluation agenda for the agency, coordinating research design with program goals, marshalling the necessary resources for solid studies, and creating an expectation of program accountability. Follow-on activities include quality reviews of evaluation products, disseminating evaluation finding, to a way to integrate evaluation results into subsequent program or funding decisions. Combined with evaluations that are rigorously designed and executed, these components make up an evaluation system that can provide departments with vital evidence on programs. Although the study revealed no states with such comprehensive evaluation practices, many states had commendable components of such a system.

States have several options to escalate their ability to critically analyze the performance of their education programs. For states with large education agencies, growing an in-house operation of qualified and experienced staff with appropriate tools and resources can yield significant net benefits in just a few years. As a thought exercise, consider a state education department budget is $100 million. Assume a level of spending of $1 million a year on in-house evaluation activities. If the evaluations lead to retargeting just five percent of resources after five years, the enterprise will be net positive. With less conservative parameters, it is clear that the break-even point would occur sooner and have a greater impact over time.

Smaller states may not be able to support internal full-service research and evaluation offices, but they can still take steps to increase their evaluation activities. The opportunity exists to collaborate across states to examine common programs. For example, two or three small states might jointly secure outside assistance to design templates for evaluations of common policies, such as charter schools, professional development, or reading initiatives. Another possibility would be to pool resources to conduct shared evaluations using the same study elements but separate data collections. Significant savings to participating states could be realized. Also, collaborative staff training in research and evaluation could effectively address the deficit mentioned by nearly all the states. Finally, states could cooperate to provide peer review of products, using staff from a neighboring state to check the quality and accuracy of work. This alternative is cost-effective and reinforces the value of regular evaluation practice.

The U.S. Department of Education can take a number of actions to influence the pace of growth. Already, the Department has established a high standard of quality for the evaluation designs it solicits under various funding programs. This important step could be expanded. The Department can extend its current leadership on measuring student outcomes to include the expectation that states regularly will relate those outcomes to various school program choices. And while NCLB leaves the choice of outcome measurement in the hands of the states, the Department can provide valuable guidance on the most useful measures for states to collect to satisfy both state accountability requirements and more target program evaluations. For example, the Department might assist states to define a larger set of outcomes, or to refine the measures they already collect to produce more consistent and reliable variables.

As a related matter the Department has a pivotal role in articulating the opportunities and limits that arise from the Federal Education Research and Privacy Act (FERPA). The development of student level data on education performance creates new opportunities to increase the rigor of evaluation studies. Accordingly, a new balance must be struck between the important rights to privacy and the potential gains from improved knowledge about education program performance. The Department can be helpful in establishing that new equilibrium.

Possibly in conjunction with Education Commission of the States or the Education Leaders Council, the Department could sponsor workshops on pertinent research and evaluation topics. For example, states could benefit from a session covering the

expanded role of experimental designs in evaluation research and federal program funding. Through its funding of the Regional Education Laboratories, it could require technical assistance to states in the design, execution, or interpretation of robust evaluations. Where common programs operate with a region, opportunities exist to facilitate shared evaluation.[11]

With deeper involvement, the U.S. Department of Education could foster state-federal partnerships to incubate research and evaluation activities in states that currently do not perform such work, or to expand the current levels of activities in states that are currently active. Several mechanisms are possible, but the central objective of such a strategy would be to help the state education agencies target one key area of evaluation practice, such as study design or original data collection, and build their internal capacities in that area. In this way, states could self-select the component that most closely fits their current needs.

Both state and federal education agencies recognize the need for more and higher quality evaluations of education programs. This report has identified both the shortcomings of existing state operations and steps that states can take to begin improvements. A consistent and continuing commitment will be needed to create the volume and quality of

[11] Many of these suggestions were shared coincidentally by education policy makers, suggesting a receptive audience to any offerings. See discussion of Gary Huang, Mindy Reiser, Albert Parker, Judith Muniec, and Sameena Salvucci in the following report: Ralph, John, Project Officer. *Institute of Education Sciences Interviews with Education Policy Makers.* Arlington: Synectics for Management Decisions, Inc., 29 January 2003.

program evaluation that can positively influence the availability of effective education programs and, ultimately, the educational achievement of America's students.