DOCUMENT RESUME

ED 465 851 UD 035 132

AUTHOR Levin, Henry M.
TITLE The Cost Effectiveness of Whole School Reforms. Urban Diversity Series.
INSTITUTION ERIC Clearinghouse on Urban Education, New York, NY.; Columbia Univ., New York, NY. Inst. for Urban and Minority Education.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.; Columbia Univ., New York, NY. Teachers College.
REPORT NO Ser-114
PUB DATE 2002-05-00
NOTE 47p.
CONTRACT ED-99-CO-0035
AVAILABLE FROM ERIC Clearinghouse on Urban Education, 525 West 120th Street, Box 40, Teachers College, Columbia University, New York, NY 10027. Tel: 212-678-3433; Tel: 800-601-4012 (Toll Free); e-mail: eric-cue@columbia.edu; Web site: http://www.eric-web.tc.columbia.edu/.
PUB TYPE ERIC Publications (071) -- Reports - Descriptive (141)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Change Strategies; *Cost Effectiveness; Data Analysis; *Educational Change; Elementary Secondary Education; Program Effectiveness; Public Schools; *Research Methodology
IDENTIFIERS Reform Efforts

ABSTRACT

This report examines issues related to the cost effectiveness of whole school reform. The first section discusses the development of whole school reform models, criteria for model adoption, and challenges of whole school reform for evaluation. The second section, "Comparing Effectiveness," looks at models for evaluation (experimental, quasi-experimental, and other methods); sampling of schools (selection concerns, bias, and student population differences); school outcomes (differences in goals, values, measurement issues, evaluators, and comparability of effectiveness reports). The third section, "Comparing Cost Data," discusses cost methodology (general principles, early cost study experience, consideration of resource reallocation, and cost recovery). The fourth section, "Conclusions and Recommendations," examines six issues that should be considered to obtain comparability for cost effectiveness purposes. (Contains 66 references.) (SM)

HENRY M. LEVIN

THE COST EFFECTIVENESS OF WHOLE SCHOOL REFORMS

The Cost Effectiveness of Whole School Reforms

Henry M. Levin
Teachers College,
Columbia University

URBAN DIVERSITY SERIES NO. 114
ERIC CLEARINGHOUSE ON URBAN EDUCATION
INSTITUTE FOR URBAN AND MINORITY EDUCATION
MAY 2002

ERIC Clearinghouse on Urban Education

525 West 120th Street
Box 40
Teachers College, Columbia University
New York, New York 10027
212/678-3433
800/601-4868
Fax: 212/678-4012
Email: eric-cue@columbia.edu
World Wide Web Site: http://eric-web.tc.columbia.edu/

Director: Erwin Flaxman
Associate Director: James M. Lonergan
Managing Editor: Wendy Schwartz

This publication was produced by the ERIC Clearinghouse on Urban Education with funding from the Office of Educational Research and Improvement (OERI), United States Department of Education, under contract number ED-99-CO-0035. Additional support was provided by Teachers College, Columbia University. The opinions expressed in this publication do not necessarily reflect the position or policies of OERI, the Department of Education, or Teachers College.

Table of Contents

Acknowledgements

This is a revised version of "Issues in Designing Cost-Effectiveness Comparisons of Whole-School Reforms," a chapter in *Cost-Effectiveness and Educational Policy*, the 2002 Yearbook of the American Education Finance Association, edited by H.M. Levin and P. McEwan. The author wishes to thank Tom Cook, Gene Glass, and Howard Bloom for reviewing the study. He is especially grateful to Steve Barnett for extensive comments.

Introduction

The Development of Whole School Reform Models

The deep concerns about U.S. education in general, and particularly the education of students in at-risk situations, have led to searches for comprehensive new models of school reform. Previous attempts at reform focused on innovations in particular parts of a school, such as curriculum, instructional strategies, organization, staff development, use of educational technology, and so on. What has distinguished the new breed of school reform has been its emphasis on transforming the school in its entirety, including all of the above dimensions and more—what is known as whole school or comprehensive school reform. Much of this focus has been on changing the culture of the school, the beliefs, expectations, and images of what are appropriate educator, family, and student roles, at the same time as instituting new instructional practices that promise stronger educational results (Finnan & Levin, 2000).

Until about 1980 the traditional approach to improving schools was to identify school challenges individually and address them idiosyncratically. Most U.S. schools adopted new curriculum packages in different subjects, technology infusions, reductions in class size, new approaches to instruction such as cooperative learning or project learning, new organizational innovations such as block scheduling, and so on. These interventions were typically done on a piecemeal basis as problems were identified that required a response. Some schools experienced five or more different "reforms" in a single year and many times this number over a decade. The more fundamental features of the school typically remained intact as reforms were simply grafted onto existing institutions and their dominant practices. Over time, schools would give up on specific reforms and replace them with others, often inserting each superficially into a school environment that was unreceptive. In most cases there was little change in the long term as these individual reforms failed to modify school operations in any substantial way (Cuban, 1993).

In the 1980s the focus of reform began to shift from idiosyncratic and piecemeal attempts to addressing the school as a whole. This shift was largely galvanized by the work of Ron Edmonds

(1979) which attempted to capitalize on ways in which effective and ineffective schools differed. Reformers such as James Comer, Ted Sizer, Larry Lezotte, Carl Glickman, Robert Slavin, myself, and others developed models for school change that addressed the entire school by attempting to alter the organization of schools, the use of resources, decision making, instructional strategies, and information flows (Levin, 1997). This became known as whole school reform or comprehensive school reform. By the 1990s it had become a prominent trend, with the New American Schools Development Corporation (NASDC) seeking "break the mold" models for schools, the publication of major works on whole school change (e.g., Fullan, 1991; Hargreaves, Lieberman, Fullan, & Hopkins, 1998), and the establishment of Federal legislation for Comprehensive School Reform Development (CSRD).

By the year 2000 there were dozens of such whole school reform models. Among the best known are the Coalition of Essential Schools created by Theodore Sizer, the School Development Program by James Comer, the Success for All endeavor by Robert Slavin, and the Accelerated Schools Project by me. In addition, there are the Core Knowledge Project of E.D. Hirsch, the Padaeia Project of Mortimer Adler, the Effective Schools movement of Larry Lezotte, and seven initial projects of the New American Schools Development Corporation (now called New American Schools [NAS] with an expansion to more projects). Even this list is far from complete, but it provides a bewildering array of choices available to school districts and schools that wish to remake education.[1]

It is important that I stipulate that I was the founder and director of one of the school reform models that is referred to, the Accelerated Schools Project. I should add that I admire all of the reform efforts referred to in this paper, although each is based on different premises on the purpose of and strategy for school change.

Criteria for Model Adoption

The central question that arises is how school systems can choose among these different approaches. One general policy tool for comparing social interventions is to evaluate them according to their costs and effects or costs and benefits. Cost-effectiveness analysis compares interventions with common goals to ascertain which

have the strongest results relative to their costs. Cost-benefit analysis compares interventions with similar or different goals to see which have the greatest benefits relative to costs, under the assumption that benefits can be measured monetarily. In principal, comprehensive school reforms could be evaluated for their costs and effectiveness or costs and benefits and compared with each other to see which ones are most promising. In reality, some may work better in certain contexts and for certain populations than others, so such comparisons would need to be made for particular settings. But, the overall notion of applying a policy tool like cost-effectiveness analysis to school reforms would seem to be a high priority.

The initial goal of this monograph was to construct cost-effectiveness comparisons of some of the existing whole school interventions on the basis of available data. For reasons that will become clear, existing data are not adequate to make these comparisons. In short, school reform evaluations differ so much in evaluation methods, sampling, measurement of outputs, and interpretations that the results are not scientifically valid for making objective comparisons, despite comparative claims on behalf of different models. Further, the costs of replicating the school reforms have not been estimated in a consistent or defensible manner. With few exceptions, available cost data are inconsistent, incomplete, rarely based upon careful and systematic methods, and miscalculate reallocation of resources. Accordingly, this paper is devoted to the issues surrounding cost-effectiveness studies of whole school reform rather than to a comparative analysis of results. It is hoped that the discussion and guidelines will ultimately lead to valid cost effectiveness comparisons.

Challenges of Whole School Reform for Evaluation

To understand the challenges that comprehensive school reform has raised for cost effectiveness analysis, it is necessary to visit briefly the previous studies of cost-effectiveness analyses in education. Cost-effectiveness comparisons require that the alternatives being considered have common objectives so that their results can be readily compared. Costs also need to be measured in a uniform way, relying on the ingredients or resource method (Levin & McEwan, 2001). With common metrics for the cost and effectiveness components, it is possible to compare cost effectiveness across

alternatives for achieving the same objectives.

By adding modest interventions such as a new curriculum or instructional approach at a cost of less than $100 or so per student out of a total school budget of $5-10,000 per student, the task is simplified. That is, the overall school is left intact, and it is only necessary to isolate costs and effectiveness of the specific change. Costs are usually an add-on that can be identified by stipulating the additional resources that are needed. Effectiveness can be measured by the changes in results that are induced by the intervention.

Interventions that were evaluated comprised programs that were added to existing schools, such as computer-assisted instruction, a different curriculum in a specific subject, smaller classes, longer school days, peer tutoring, and so on (Levin, 1991; Levin, Glass, & Meister, 1987). Measures of effectiveness explored whether each of these types of interventions, when added to a regular school program, had an impact on student achievement and the magnitude of that impact. Cost measures examined only the marginal or additional costs of these interventions to the school, not the overall costs of school operations. Results were converted into units of effectiveness for a given cost and compared across alternatives.

In these cases the intervention could be readily identified as an "add-on" to the school program, and its costs and effectiveness could be measured somewhat independently of the existing program. Of course, some "add-ons" might work better at some sites with some types of students and existing programs than at other sites. Such differential effectiveness could be taken account of by looking for statistical interactions between site-specific variables and the effectiveness of an intervention (Summers & Wolfe, 1977) or by carrying out the analyses separately for different populations or contexts (Grissmer, 2002). At the same time it was relatively easy to separate out the added ingredients or resources that a school needed to implement each of the programs. And effectiveness could be limited to one or two specific program outcomes when the interventions were compared, e.g., reading programs (Levin & McEwan, 2001).

But with whole school reform there is a transformation from a traditional school to a restructured one, with a potential impact on all the goals of the school, not on just one or two. Both reallocations of existing resources and added resources are pertinent to whole school reform. These aspects of whole school reform create an enor-

mously greater challenge in doing a cost effectiveness (or even just an effectiveness) comparison among different programs. Indeed, it is these types of methodological problems that are the subject of this paper.

A related challenge is that a focus on a single output such as reading or mathematics competency (or even both considered together) is an inadequate basis for considering the productivity of a school. Any formulation that considers only these outputs will fall prey to considering only a portion of outcomes that the schools produce. It will mean that major outputs will be unaccounted for. Thus, a comparison between two whole school interventions that focuses on only a single output or dual outputs will not monitor what is happening to other outputs. Certainly, by shifting resources from the production of unmeasured outputs to measured ones, it is possible to obtain more of the measured ones. But, this is only a partial measure of the total output of the school as it would be for any multi-product firm.

Consider that even a short list of what schools are expected to produce is formidable. It would include raising students' proficiencies in many subjects including reading, writing, speaking, mathematics, science, social studies, art, and physical education, as well as their acquisition of a large number of social values and behaviors. With respect to the latter, schools emphasize working cooperatively with others, following rules, accepting constructive criticism, planning a project, setting goals, seeking out necessary information, resolving conflict appropriately, and respecting differing viewpoints, to name just a few that come to mind. Inkeles (1966) suggests still more social skills that schools are expected to provide to create competent adults.

Given the multiple teaching tasks of schools, it is possible to increase one output without improving productivity by neglecting other outputs. For example, if resources that were previously devoted to other outputs are focused more intensively on mathematics, it is possible to improve mathematics achievement at the expense of results in other areas of learning. Allocating more of the school's personnel and greater instructional time to mathematics, at the expense of other subjects and social behaviors and attitudes, can improve mathematics achievement. As long as all the outcomes of a school are monitored, this shift in resource allocation will be reflect-

ed in the rise in mathematics achievement and a measurable reduction in the attainment of other school goals.

But, as we will see below, many studies measure only one or two school objectives, such as reading or mathematics or graduation rates, rather than the plethora of outcomes that schools are expected to produce. The result is that it may be impossible to determine what is being sacrificed among other outputs to obtain given improvements in the output under scrutiny, although, surely, some area of learning is shortchanged by the shift in resources. As an example, Bowles and Gintis (2000) provide cogent evidence that non-cognitive aspects of schooling which are not even measured in school accountability systems may be the dominant determinants of education's effect on earnings, rather than cognitive aspects measured by tests. Further, over the years a number of subjects such as geography have disappeared from U.S. schools. Does it matter if most Americans confuse Australia with Austria or do not know the continents or locations of nations in an age of globalization and international conflict? A similar issue is evident in high stakes testing where teaching efforts, curriculum, and test preparation shift to what is tested from what is not (McNeil & Valenzuela, 2000).

To the degree that whole school reform is undertaken largely with existing resources, when evaluations are based upon a single objective or narrow range of objectives, resources are likely to be reallocated from existing uses to those most closely aligned with those objectives. Thus, unless there is a way to assess the impact of the reform on all outcomes—or at least on all major outcomes—any attempt to limit effectiveness studies to a single objective will be suspect as an overall assessment of effectiveness of the reform model. As we will see below, this also raises challenges for the measurement of costs, because reallocations of an existing budget may not be costless in an economic sense. The next two sections review the implications of this background discussion for measuring both effectiveness and costs.

Comparing Effectiveness

Ideally, we would like to obtain comparable data on school effectiveness to compare among school reform models. Such data could be combined with comparable data on costs to ascertain the cost effectiveness of each of the reform approaches. Slavin and Fashola (1998) and Herman (1999) have published comparisons of effectiveness on what they assert is a review of "evidence."[2] In this section, I will argue that such comparisons—in the absence of methodological, sampling, and other adjustments—are invalid. The lack of comparison validity is not due to subtle issues. It is due to fundamental differences among designs and procedures that can account for different evaluation outcomes beyond differences in the impacts of the models themselves (see Hunter & Schmidt, 1994, for an overview of some of these issues). Each of these differences in treatment of the studies will be addressed below.

In particular I will address four issues that must be considered in doing a comparative cost effectiveness analysis. The first is the question of whether the overall evaluation approach is valid and how the particular choice of evaluation method may affect the magnitude of reported effectiveness. Second is the question of how representative the schools sampled for evaluations are. Third is the issue of multiple education goals and how they will be accounted for. Fourth is the matter of potential bias in evaluations done by school reform sponsors or their representatives relative to independent evaluations. Each will be taken in turn.

Models for Evaluation

A major concern is whether the evaluation model is an adequate one by which to obtain valid results. There are two primary approaches used in the literature and an eclectic third strategy.

(1) Experimental: The pure experimental model requires schools to be randomly assigned to treatment and control groups (Boruch, 1997). If the treatment and control groups are large enough, comparability in school features is assured so that any difference between the two groups after implementation of the treatment could be attributed to the treatment. This type of approach is not easily applied to whole

school reform, since all of the reforms require that schools select the specific reform that is adopted (typically approval by 80 percent or more of the teachers) rather than permit assignment of the reform to the school. In some cases, however, this process of "informed consent" is breached as specific reforms are pushed on schools by school districts (Datnow, 2000). Still, all of the major comprehensive school reforms attempt to ensure an active process of informed choice and "buy-in." Once schools select or buy in to a specific reform, they are usually accepted by the sponsor of the reform.

The only experimental studies that I could find were those undertaken by Thomas Cook and his colleagues. Cook, Hunt, and Murphy (2000) compared ten inner-city Chicago schools using the School Development Program or Comer model with nine comparison schools over a four-year period.[3] The Comer schools and the comparison schools were selected randomly from a population of low-achieving Chicago schools, all that had volunteered to adopt the reform. The evaluators found small achievement advantages for the Comer schools relative to the control schools (about three percent over three to four years) as well as advantages in student behavior and attitudes. Cook et al. (1999) also used the experimental methodology to study academic and other outcomes in a randomized study of 23 middle schools in an urban county in Maryland. Differences were found in favor of Comer schools in psychological and social outcomes, but not student achievement.

(2) Quasi-Experiments: Quasi-experiments represent attempts to emulate experimental conditions as closely as possible, in the absence of random assignment. Almost all evaluations of whole school reforms fit this category. But, as Cook and Campbell (1979a; 1979b) point out, there are many threats to the validity of such evaluations, so the reader should be aware that although certain results are claimed in such an evaluation (even what appears to be a sophisticated one) they may not be substantiated. The typical quasi-experimental design attempts to compare schools receiving the intervention with similar schools that are not receiving the intervention. Statistical adjustments are often used to attempt to adjust for differences between intervention and comparison groups which could affect outcomes.

Robert Slavin and his colleagues have carried out a large number of studies where intervention schools and matched comparison

schools are assigned directly by the evaluators rather than randomly (Slavin & Madden, 1999; 2000). They conclude that their Success for All and the Roots and Wings models show very substantial gains in student achievement relative to the comparison schools. The Center for Policy Research in Education (CPRE) evaluated schools from the America's Choice model in three cities, using as comparison schools those that had not adopted the model in these cities (Supovitz, Poglinco, & Snyder, 2001). Statistical controls were provided for demographic and other factors that might affect achievement. The authors found small achievement advantages for the America's Choice schools at most grade levels for the first year of implementation. Ross, Wang, Sanders, Wright, and Stringfield (1999) compared achievement gains in a large number of schools undertaking whole school reform in Memphis with a "matched" group of schools not undertaking the reforms. They also compared the achievement effects of the different reform models and found substantial achievement gains of those schools participating in the whole school reforms relative to the comparison schools, using a value-added model (see Sanders and Horn [1995] for a presentation of the value-added model).

Millsap et al. (2000) found no difference in achievement between 12 Comer schools in Detroit and a set of matched, comparison schools, although they did find that those schools with the best implementation of the reform had better achievement than the comparison group. In a similar type of study no overall difference in achievement was found between 12 Core Knowledge Schools and matched comparison schools, but an effect was observed for Core Knowledge Schools with a high level of implementation (Stringfield, Datnow, Borman, & Rachuba, 1998). A study of five Core Knowledge Schools in Maryland found mixed results in comparing achievement in schools with the intervention and comparison schools (Mac Iver, Stringfield, & McHugh, 2000).

A different quasi-experimental design is that of interrupted or discontinuous time series (McCain & McCleary, 1979). Here, the pattern of achievement for a particular sample of schools is evaluated over time, for several years prior to the reform and then several years following it, to test whether the pattern was altered following the intervention. Statistical adjustments are made for other changes in the school over that period, such as changes in socioeconomic or racial composition of students. The Manpower Development Research

Corporation (MDRC) used this technique to evaluate third grade achievement for a national sample of eight Accelerated Schools with about 3,000 students in the study (Bloom et al., 2001; Doolittle, 2001). The improvement in mathematics and reading achievement by the fifth year of implementation was 7-8 percentiles.

(3) Other Methods: Several other methods have been used to evaluate whole school reforms. The RAND Corporation has examined the degree to which such schools have outpaced the average achievement gains of the districts where they are situated (Berends et al., 2001). This is a fairly common approach in which the district is viewed as the comparison standard where all schools in the district are assumed to be subject to the same non-reform influences. Correlational approaches have also been used in which statistical models are designed to isolate the effects of a reform intervention from other factors which may influence achievement, such as school resources and student characteristics.

Perhaps the weakest design is that of year-to-year achievement gains for individual schools without comparison data. This is the typical format used by school districts to report progress. Even with gains of reforming schools, the question is whether those gains are greater than for comparable non-reforming schools. An equally serious challenge is the long-term reliability of such short-term, measured gains. Recent studies have found that a high proportion of achievement change from one year to the next is transitory and due to idiosyncratic factors (Kane & Staiger, 2001). This means that gains from year-to-year may not be permanent, but due to temporary circumstances, such as an especially strong or weak student cohort or a disruptive period when testing was done. Kane and Staiger estimate that less than half of the average achievement gain in reading between fourth and fifth grades (the grades for which they tested the relationship) showed persistent differences between schools.

What is noteworthy is how many different methodologies are used and the variants of each in terms of the sampling and measurement factors discussed below. Even within these evaluation models, substantial differences in their implementation can affect results. As noted, the choice of comparison schools in quasi-experimental studies is often arbitrary. With the exception of the studies by Cook and his colleagues (1999; 2000), comparison schools are not randomly

chosen. For example, Supovitz et al. (2001) chose as comparison schools those schools not carrying out the America's Choice reform in the three districts that were studied. Evaluations of Success for All (Slavin & Madden, 1999) provide few specific details on how comparison schools are chosen other than an attempt to provide a demographic match.

Since all of the major whole school reforms require buy-in with support from 80 percent or more of their teachers or school staff, it is likely that those with stronger leadership and committed staff will undertake the reform. *Prima facie* this suggests that the evaluator is comparing energized schools ready to undertake reform with schools that are not, rather than comparing schools that are comparable in every way except for the nature of their programs. This fact is likely to lead to overstatement of the measured effectiveness of the reform.[4] But, differences in the rigor of the buy-in requirements among models will create differences in this bias, with the models that demand greatest commitment in buy-in creating greater selection bias in favor of success. At the same time, the standards for choosing comparison schools may differ substantially, resulting in estimated effects conditioned upon the readiness, leadership, and enthusiasm of schools undertaking reform versus the lassitude of the comparison school determining outcomes, rather than the impacts of the reform models themselves.[5]

Sampling of Schools

Selection Concerns

One of the great challenges in education is replication. Even when an educational intervention has been shown to have strong effects at a demonstration site, it is rare that it is replicated at other sites with similar results. Indeed, the history of educational reform is more a testimonial to constancy and resistance to change than to change itself (Cuban, 1993; Sarason, 1982). From a policy perspective, the concern should not focus on results from experimental, demonstration, or exemplary sites, but with the potential effect of expansion of initially successful sites to new sites and scaling up from a few to many. This means that evaluations for cost effectiveness purposes should be based not upon initial results at a relatively small number of sites that have received special attention and nurturing, but on replication under the

ordinary conditions that will be found as expansion takes place. Too often educational evaluations are done in laboratory settings or in schools where university support and scrutiny are provided—factors that are unlikely to be pertinent in subsequent replications under ordinary conditions. Obviously, the first prototypes and initial replications will have the most assistance, attentive evaluations, and publicity. It is incorrect to assume that subsequent applications of the approach will be equally effective. For example, Lipsey (1999) found a very substantial difference in effect sizes in favor of demonstration programs over replication programs.

Few evaluations of educational interventions can be found that reflect what happens under the most routine replications. It is widely recognized that published evaluations overstate the average or typical effects of interventions because poorly performing sites will not be evaluated and evaluations showing poor results will not be reported or published (Begg, 1994; Glass, McGaw, & Smith, 1981). Some authors even recommend eliminating from consideration evaluations of those sites that do not implement a model "correctly." In fact, Slavin and Madden (2000) suggest that virtually all instances of poor evaluation results for the Success for All model are a consequence of poor implementation rather than flaws in the model. But this position raises questions of whether the implementation process accompanying a reform is an integral part of it or is independent of it. Clearly the truth lies somewhere between these two extremes.

The RAND Corporation analysis of New American Schools attempted to measure the degree of implementation of different models rather than an absolute measure (Berends et al., 2001). Surprisingly, the RAND study found no linkage between the degree of implementation and school performance among the 163 New American Schools for which data were available. If a decision maker asks the question, "What is the expected effectiveness of a particular model under 'typical' conditions?", the probability of poor implementation should be included in the overall assessment of effects. Suchman (1971) has provided one of the best conceptual discussions distinguishing between the failure of theory and the failure of program implementation. Sites that implement the models well are not typical of the average implementation, and high implementation may also be related to strong leadership, staff comaraderie, and staff talent—factors that are independent of the model. In many cases, if not most, it may be impossi-

ble to know if a chosen school will provide good implementation. Therefore, the decision maker is more likely to ask the question, "If schools adopt a particular model, what is the likely outcome?" (not if they adopt a likely model and succeed at implementation).

Thus, a first concern with respect to a cost-effectiveness evaluation is whether the study of effectiveness has been done on a typical replication that is reproducible from site-to-site under "ordinary" circumstances. Obviously, the most appropriate estimate of effectiveness would be to evaluate the full population of replicated sites or a representative sample of adequate size. As will be shown below, results may be radically different from study to study of the same school reform models, at least in part from sample selection.

Bias

There are at least three issues with regard to sampling bias. First is the bias mentioned above: that only the most energized schools anxious to change will buy in to the reforms, making them unrepresentative of other schools that might be similar in demography, size, and location. This bias boosts the apparent effectiveness of all of the models requiring buy-in. But differences in the rigor and verification process of adoption criteria will create differences in the degree of bias among the different school models. Those models that set and enforce the most stringent criteria for staff participation in the adoption process will likely have schools that are more highly motivated and prepared to implement reforms than those that do not, independently of whatever reform is being implemented (Datnow, 2000).

Second, there is a question of whether the schools in the evaluation samples are representative of all schools participating in the reforms, of only those with good implementation, or of samples of convenience (where data were available or results were promising). Only Cook and colleagues (1999; 2000) have made an attempt at random assignment of schools, and only within two specific localities, even though the model that they evaluated was implemented nationally. Although the MDRC study of Accelerated Schools (Bloom et al., 2001; Doolittle, 2001) represents an attempt to study a national sample of implementing schools with eight years of data, data availability itself affected the nature of the sample. The recent study of America's Choice schools is based upon schools in the three school districts with

relatively large adoption of its schools, suggesting strong district support for its model relative to the more typical situation where district adoption of a reform model is not widespread among schools. Although there are many more evaluations of Success for All than of other models, there is no information suggesting that an attempt was made to select representative sites.

These sampling problems undermine attempts to predict the effectiveness of each reform model for future schools that might consider adoption. The buy-in requirements mean that the results cannot be extrapolated to other schools that have not gone through a similar buy-in. The bias towards evaluations covering only schools with high implementation tends to overstate results for more typical situations. Also, it is likely that schools selected for evaluation are those benefiting from favorable conditions such as strong district support, and this bias may differ from study to study. This selection bias means that even if the evaluation models used in all studies were similar, differences in sampling would threaten valid comparisons. Thus, the overall results across studies cannot be generalized to the overall population of schools, even those with similar demographics or locations. It is noteworthy that the RAND study of comprehensive school reform models sponsored by New American Schools found that slightly fewer than half of the reforming schools had achievement gains greater than the average gains for their districts, about what chance would predict (Berends et al., 2001). All of the models report far greater success in their own evaluations based—in part—on different samples from those used by RAND.

Student Population Differences

A third way in which sampling varies is among the types of students who are evaluated. Different studies eliminate different types of students from the testing base, such as those with learning disabilities or in bilingual programs. Typically, evaluations of whole school reforms focus only on those students who were in the school continuously over the evaluation period. But, the most educationally needy and disadvantaged students have high mobility rates and are found in schools with high student turnover. For example, Kerbow (1996) found that in the typical Chicago elementary school only half of the students were still enrolled at the school after three

years. Limiting the evaluations to students who remain at a single school site restricts the evaluation to those students in stable situations who generally have higher achievement and achievement gains, irrespective of the reform model (Kerbow, 1996; Rumberger 2001). Further, such children have an additional advantage because they are exposed to the reform for the entire duration of the period of evaluation. In contrast, students with high mobility spend less time in a particular school and have less exposure to the reform, often being churned among many different schools over a one- or two-year period. At the very least, studies that eliminate mobile students should report that their results apply only to stable students who have had continuous exposure to the reforms, not to all students attending the schools that are evaluated. Unfortunately, such studies tend to generalize their interpretations to all students.

But, in addition, the variation in the treatment of student mobility among studies contributes to their non-comparability. For example, evaluations of Success for All have largely been limited to those students who attend the same school continuously for all of the years in which the school is evaluated, as many as five years. Such studies can say little about the effects of the reform on the many students—typically with lower achievement levels and lower achievement growth—who move to other schools (Kerbow, 1996). The studies by Cook et al. (1999; 2000) and Millsap et al. (2000) of Comer schools and by Stringfield, Datnow, et al. (1998) of Core Knowledge Schools also restrict themselves to students attending their schools for the duration of the program treatment and evaluation. In contrast, the evaluation of Accelerated Schools (Bloom et al., 2001; Doolittle, 2001) includes all students in the school at the time of the third grade testing. It was estimated that about half had not received three years of exposure to the Accelerated Schools model and one-third had not even received two years of exposure (Bloom et al., 2001). Obviously, a comparison of stable students who have been exposed continuously to a reform with one that also includes students who have not had that advantage will bias the result in favor of the reforms that eliminate mobile students from the evaluation. The evaluation of America's Choice by Supovitz et al. (2001) takes a middle position by including all students in the two districts for which there were records over a one year duration. That is, students who moved to other schools in the district during that year were included, but not those who moved to or from other districts. Studies that limit their

analysis only to students who were attending the school for the duration of the evaluation period will show higher achievement effects than those that include all students in the analysis. This difference in student sampling undermines comparability of results.

School Outcomes

Differences in Goals

A particularly difficult evaluation challenge is that of the multiple and different outcomes desired by the various school reform models. For example, the School Development Program, the Coalition of Essential Schools, Different Ways of Knowing, and the Accelerated Schools Project all require schools to set their own priorities and address them. Schools may focus on improving achievement in different subjects and on increasing student involvement in projects, and may establish a range of other goals that are likely to vary from school to school. The School Development Program strives for the integration of families into the school community, while the Accelerated Schools Project and the Coalition for Essential Schools develop a learning community with a specific philosophy based upon constructivist learning theory. Different Ways of Knowing utilizes the arts to connect all subjects. In contrast, Reading Recovery and Success for All concentrate on early childhood reading proficiency by using explicit strategies and materials. Core Knowledge emphasizes mastery of a specific knowledge base and use of a particular curriculum.

There is absolutely nothing wrong with a schoolwide strategy to improve student proficiency in a single subject. However, it is inappropriate to compare effectiveness in only one or two subjects among different school reform models, when some focus only on that subject and others focus more broadly on a variety of outcomes. And indeed, this is the dilemma. Schools are multi-product firms with multiple goals. School reforms that focus on only one dimension can show superior results by concentrating their efforts on that goal. But an evaluation on a single dimension is biased against those reform models concerned with improvements in multiple outcomes. Such a comparison also violates the tenet in cost-effectiveness analysis that only interventions with common goals should be directly compared—and all major goals must be included in the evaluations.

Differences in Value

The case of multiple outputs also raises the question of how to value effectiveness. For example, if one model does better on reading and the other on mathematics in the simplest case of two outputs, how is comparative effectiveness of the two alternatives determined? The standard approach is to place values on each of the two outputs using utility scales or other devices (Levin & McEwan, 2001). Of course, an evaluator could weight each of the results equally (e.g., equivalent effect sizes in each subject are given equivalent weight). But, there is no reason *a priori* to provide equal weights to equivalent effect sizes across many subjects and other measures of student behavior. Although music and art are extremely important subjects in the curriculum, it is not clear how society should value them in comparison with reading, science, mathematics, social studies, writing, and the other core academic subjects. It is possible that music and art should be considered more important than some of these other subjects because of their intrinsic value as well as the contributions that they can make to cultivating creative talents, and, in fact, Different Ways of Knowing is a reform that uses the arts as an integrative strategy for other subjects. But what is clear from this debate over the relative importance of subjects is the complexity of the challenge of combining school results for different outcomes into an overall rating of effectiveness. There exist "solutions" to this problem, but they are highly subjective and based upon assumptions for ascertaining social value that are arbitrary (Levin & McEwan, 2001).

Differences in Measurement Instruments

How particular outcomes are measured is also an important criterion in comparing alternatives. When measuring change in academic achievement, it is common to use "effect size" as the criterion (McGaw, 1994). The advantage of using effect size is that it is a common measure that can be calculated for different tests that ostensibly measure achievement in the same curriculum subject. Thus, even though different schools and school districts utilize different testing instruments constructed by different test publishers, an evaluator can calculate the change in achievement between two periods and divide it by the standard deviation of the test to get an effect size. Presumably the effect size

is a common metric that allows comparison of effectiveness.[6]

However, each test instrument measures its domain in different ways. For example, fourth grade mathematics tests may differ in the weight given different types of mathematics operations, applications, concepts, and word problems. Some tests will rely more heavily on math facts. At the opposite extreme, some will emphasize concepts and solution of word problems. The same students will likely score differently on different tests, depending on what is being tested and how it is tested. Thus, although it may appear intuitively plausible that test results can be compared in terms of a common metric—effect size—different tests represent different measurement systems. Some are considered to be more difficult than others because they have more complex calculations, harder problems, or rely more heavily on speedy solutions under rigorous timing restrictions. To some degree each test is measuring a different set of outcomes, and its results are not strictly comparable with the results of other tests, even though one can mechanically compute effect sizes for each.

A related problem is the issue of whether an evaluator uses a broad spectrum test or one that is tailored to the curriculum that a reform represents. If a school reform model uses a prescriptive approach to curriculum content and instructional strategies, it will seem natural to align it with a specific test that measures the success of that approach. Thus, the criterion of effectiveness will match the intervention closely. But where the school reform model leaves a wide range of discretion for schools to construct their own curriculum content and instructional strategies, it is unlikely that one test instrument will closely match the goals of all the schools using that model. Further, school districts and states mandate many different tests, none of which may align perfectly with the subject area as defined by the various school reform models.

The result is that achievement measured by a test that is aligned with the specific curriculum content that is taught and the instructional strategies that are used is likely to be greater than by an independently prepared test. In general, this means that school reforms using aligned tests to measure results will show better results than those using more general testing systems. As an example, a Success for All school was matched with a comparison school to assess the gain in reading over an academic year. On the aligned tests used by Success for All, there was a statistically significant difference in reading in favor of

the Success for All school. But when the Tennessee Assessment System was used to make the comparisons, there were no statistically significant differences between the Success for All and the comparison school in reading or any other subject (Ross & Smith, 1994).

Several of the school reform models—such as the Coalition of Essential Schools and Accelerated Schools—place an emphasis on student performance on real-world or authentic tasks rather than on test problems which are usually far removed from those tasks. For example, instead of a test of knowledge of chemistry, a performance evaluation might require the student to analyze a "mystery" substance for its elemental components, testing the student's abilities to use logic, intuition, knowledge, and laboratory procedures to accomplish a real-world challenge. An evaluation of a student's mastery of a Shakespearean tragedy might entail the writing of a short work embodying the style of that form of literature with a presentation before the class and a demonstration of how it meets the criteria, as well as knowledgeable responses to questioning.

Consider the comparison for effectiveness of a school reform that is focused on performance assessment with one focused on criteria that are directly measurable by a standardized test. The standardized test will be far more closely aligned with the latter than the former. And what appears to be a neutral measure of assessment, such as a chemistry test or English test, or a history test, or mathematics, test will not be neutral at all. For the test-driven curriculum, students will be repeatedly tested in the standardized format with both testing experience and curriculum goals contributing to their test performance (McNeil & Valenzuela, 2000). In the school emphasizing performance assessment, both the goals and the students' experiences will not match up well to standardized tests. The result is that comparisons of effectiveness of school reforms will be problematic when a single set of standardized tests is used to assess reforms whose goals differ so substantially from reform to reform.

Differences in Evaluators

A particularly challenging aspect of the existing evaluations of effectiveness of whole school reforms is that most have been done by the sponsors of the reforms. Almost three decades ago James Q. Wilson (1973) set out two laws that he believed apply to all cases of social sci-

ence evaluation of public policy:

> First Law: All policy interventions in social problems produce the intended effect—if the research is carried out by those implementing the policy or their friends.
>
> Second Law: No policy intervention in social problems produces the intended effect—if the research is carried out by independent third parties, especially those skeptical of the policy. (p. 133)

Wilson is not accusing developers or their friends of fudging the data to support effectiveness claims, but simply stating that different standards of evidence and method used by evaluators who are assessing their own interventions tend to impart an upward bias. There are many areas for judgment calls in evaluation. In general, the sponsors or their colleagues accept conditions, methods, and measures that are more favorable to their interventions than they would if they were evaluating competing models. Scriven (1976) has written comprehensively and perceptively on evaluation bias.

Third-party evaluations are defined as those meeting two conditions. First, the evaluations are carried out by independent evaluators who have virtually no personal or institutional links to the sponsors of the reform. Second, they are not funded directly by the reform sponsors or developers. Using these criteria, Cook, Hunt, and Murphy (2000), Cook et al. (1999), and Millsap et al. (2000) have carried out third-party studies of the Comer model. MDRC has carried out a third-party study of Accelerated Schools (Doolittle, 2001); the Center for Social Organization of Schools at Johns Hopkins University (Mac Iver et al., 2000; Stringfield, Datnow, et al. 1998) completed third-party evaluations of Core Knowledge Schools; and Supovitz and colleagues (2001) at the Center for Policy Research in Education undertook a third-party evaluation of America's Choice. (The study of America's Choice was funded by the sponsoring organization, so it does not meet the condition of independent funding.) The RAND Corporation has carried out third-party evaluations of the New American Schools models.[7]

Most evaluations of whole school reforms have been carried out by the sponsors of the reforms rather than by third parties. Even

attempts to summarize results across the different reforms have been carried out primarily by those associated with specific reforms. For example, a highly-publicized, "third-party" review of evidence on effectiveness of school reform models was carried out under the aegis of the American Institutes of Research (Herman, 1999), but the director of the study had recently shifted employment from an organization sponsoring one of the reforms and had collaborated previously in a laudatory evaluation of that reform (Stringfield, Millsap, & Herman, 1998). Studies undertaken by school districts or their schools may be biased in either direction depending upon their point of view on a reform. This certainly seems to be a point of contention surrounding the evaluation of the whole school reforms in Memphis, which resulted in a report prepared for its school board recommending the abandonment of all such reforms (Calaway, 2001).

As the literature predicts, third-party evaluations tend to find more modest effectiveness results than the assessment results of studies of the sponsors of those reforms.[8] The RAND evaluation of New American Schools found that fewer than half of the schools showed achievement gains in reading or mathematics that were greater than those of the districts overall in which they were located. Bear in mind that by chance about half will be above the district average. The contrast between the assessments reported by the sponsors of the models and the RAND result is striking. For example, the summary of evaluation results for Success for All and Roots and Wings (Roots and Wings is an expanded Success for All model including other subjects as well as reading) reports consistent and overwhelming evidence of effectiveness, primarily on the basis of the sponsor's own evaluations (Slavin & Madden, 1999). In contrast, the RAND evaluation (Berends et al., 2001) found that fewer than half of the Roots and Wings schools had achievement gains greater than their districts. Although America's Choice showed significant gains in reading and mathematics relative to comparison schools in three school districts, the RAND evaluation of its New American Schools found that only slightly more than half did better than their districts in mathematics and only a quarter of the schools did better in reading.

In a study of achievement using the value-added approach, the first wave of reforming schools in Memphis had much better results than a matched group of non-reforming schools (Ross et al., 1999).[9] But a more recent report sponsored by the Memphis School District

that evaluated all of the schools engaged in reform, including those that started later, found virtually no improvement in achievement among the reform models (Calaway, 2001).

This contrast in results provides a good example of the dilemma of attempting to compare differences in reform models. For the same set of school reforms and the same setting, the results differ markedly. Why? The two studies differ remarkably in their samples, time frames, and methods, all matters of judgment or choice in the evaluation process. Moreover, the potential orientations of the two evaluation groups are different. The Ross et al. (1999) study, with three of its coauthors associated with one of the models and its principal author also affiliated with the New American Schools evaluations, has made evaluation choices that are more likely to favor findings of effectiveness. For example, because only the first wave of schools that volunteered for school reform is included, the sample is likely to have an upward bias relative to the later schools that were required to adopt a reform under an imposed deadline. In contrast, the study by the Memphis City Schools (Calaway, 2001) was prepared for a school board meeting in which the new Superintendent was ready to recommend the dropping of all school reforms promoted by the previous Superintendent. The result is that schools which had barely begun to engage in the reform process and which had been pressured to adopt a reform were bundled along with those that had actually implemented the reforms.

Comparability of Effectiveness Reports

In summary, evaluations designed to assess the effectiveness of the different school reform models are premised on choices of samples, measurements, methods of evaluation, and interpretations that differ markedly among evaluations. In particular, evaluations of sponsors tend to select "successful" schools and ignore those where the results are not salubrious, and to select outcome measures that are aligned with their own purposes as well as methods of analysis (e.g., in choice of comparison schools) that tend to favor their reforms. Third-party evaluations may sample differently, use outcome measures that are broader, and employ methods that are not necessarily favorable or sensitive to specific reform strategies. Among both first-party and third-party evaluations, inconsistencies in evaluation procedures from study to study are likely to account for much of the observed difference in

outcomes. A reasonable conclusion is that the body of effectiveness results that is presently available is based upon such different samples, methods, and measurements that direct comparison is inappropriate and can be very misleading. More contentious is the fact that even within a single method the evaluations of sponsors tend to overstate the effectiveness that might be expected in a random replication.

Comparing Cost Data

Two main issues emerge in considering the measurement of costs as evaluators apply cost effectiveness analysis to whole school reforms: first, how the cost methodology should be chosen and applied; and second, how reallocations should be treated. Sadly, to date, neither has the appropriate cost methodology been used, nor have reallocations of resources been treated appropriately in cost analysis of whole school reforms.

Cost Methodology

General Principles

As described above, the concept and procedures for measuring costs are fairly straightforward, certainly in comparison with those for measuring effectiveness (Levin & McEwan, 2001). The problem is that they have rarely been followed in education, and the advent of whole school reform has not changed this practice (Levin, 2001). Cost analysis begins with the recognition that resources have value in alternative use, whether paid for or donated, and the most valuable alternative use determines the cost value of the resource (Levin & McEwan, 2001). Thus, an exhaustive search must be undertaken to specify all of the resources or ingredients that are necessary for the reform model. Again, the concern is to estimate the cost of a typical replication. Often the initial implementations of a model receive considerable personnel attention and other resources as the sponsors of the intervention go to heroic measures and draw upon developmental resources to make their intervention work. Often these extra resources are not accounted for because they were thought to be incidental or not absolutely necessary to the design of the intervention. Nevertheless, they must be assumed to contribute to the effectiveness unless it can be shown that a "slimmed-down" version gets equal results. So, evaluators must be careful not to relate the effectiveness of the initial version with the costs of a slimmed-down replication. Only the ingredients associated with the particular version whose effectiveness is being assessed provide a proper basis for the cost estimate.

Drawing upon multiple sources of information will provide the

most accurate determination of required ingredients. Initially, it is best to review reports and other documents which describe the development of the intervention and its requirements. This scrutiny will sensitize the analysis to the types of ingredients that are necessary and will enable the drafting of a preliminary list that identifies the ingredients in sufficient detail so that later the costs of each can be specified. Thus, for example, specifying the need for a full-time teacher or half-time administrator is not sufficient; some details on the qualifications for these positions are also required.

Direct observation of a replication at a representative site or a random selection of sites is the second source of information. Observation can be used to verify ingredients and their descriptions. But in addition, the direct observations are used to ascertain other resources that might be used—such as personnel, facilities, equipment, materials, services, insurance—which might not have been identified or were not identifiable from the initial documents. This investigation is combined with information from a third source, that of interviews with both key personnel and those involved in daily operations. As observations are made, it is important to ask staff about their functions, what other personnel are involved (since some may be part time, occasional, and away from the site), what outside services are used, which particular facilities are necessary to the intervention, and so on. When a final list of ingredients is drawn up from the three sources (documents, observations, and interviews), it is useful to verify its accuracy with someone who has authoritative knowledge about the intervention. It is also important to determine if these are the ingredients required for replication or are partially idiosyncratic to the sites being assessed.

The ingredient information is usually arrayed on a financial spreadsheet (e.g., EXCEL) according to major categories of personnel, facilities, equipment, materials, and miscellaneous. Each of these can be divided into sub-categories. At this point it is necessary to estimate the cost values of the ingredients. One possibility is to estimate the local costs at the site or sites being scrutinized. But the problem with this approach is that such costs may be idiosyncratic to the sites and not generalizable to other sites. For example, in areas of high real estate costs, facilities may cost considerably more than in other areas. The costs of education personnel will tend to be higher in areas with a higher cost of living and higher salaries generally.

Of course, at any specific local site, costs should be estimated for that site, but when making overall cost-effectiveness comparisons, it is usually better to get a "standard cost": the average cost for a particular geographical area. These standard costs can be applied to all of the interventions that will be compared for cost effectiveness.[10]

Methods for estimating costs of ingredients are found in Levin & McEwan (2001). However, it must be kept in mind that any resources that have value in alternative uses represent a cost, even if the resource is donated. The issue of who pays for a resource is separate from whether a cost is incurred. The same is true for a subsidized ingredient such as a facility that is paid for by another level of government. The cost is the value of the facility over the life of the intervention. That and other costs can be allocated to different constituencies or entities that bear the cost (Levin & McEwan, 2001).

Costs must then be determined for each alternative and compared with the effectiveness of each. All the costs should be included in this comparison.[11] For educational interventions, the cost per student is often taken as the criterion and compared with average achievement gain per student. However, this mitigates against projects that have a large fixed cost such as those that require a substantial investment in capital equipment that can accommodate a very broad range of enrollments. At lower enrollments the cost per student will be high because the fixed costs must be divided among a very small number. However, with larger enrollments the fixed costs do not rise commensurately so that average cost per student drops. Therefore, the comparison of costs must be sensitive to different levels of scale rather than relying on a single enrollment level to estimate costs.

Early Cost Study Experience

Two early attempts to measure the costs of whole school reforms are worthy of mention. King (1994) attempted to estimate the costs of three different approaches: Success for All (Slavin et al., 1990), the School Development Program (Comer, 1988), and Accelerated Schools (Hopfenberg et al., 1993). It is important to bear in mind the different foci of these reforms: the first of these is a highly prescribed approach to preventing failure in early childhood reading, the second applies a child development process and community involvement to

the entire school program, and the third represents an effort to transform schools and classrooms by replacing remediation with educational enrichment usually provided to gifted and talented students. King used an appropriate conceptual framework in selecting the ingredients approach in which she intended to specify the resources used in each intervention and place cost values on them for purposes of comparison. However, she did not gather the ingredients directly from field implementation of the models, but relied instead on general descriptions of the models that did not provide any detail from actual experience. It would have been preferable if ingredients had been derived from the actual replications of the models, given the many replications for all three models at the time of her research, rather than from general descriptions of the models. Using the ingredients framework as a guideline, she found substantial differences in costs among the three models. Accelerated Schools had the lowest cost per student, Success for All had the highest cost, and the School Development Program was intermediate between the two. No attempt was made to measure effectiveness.

In a more recent study, Barnett (1996) attempted to compare both costs and effectiveness of the same three models. Although his study is more comprehensive than King's, he also had difficulty in providing data on costs and effectiveness. For reasons set out above about the difficulty of comparing effectiveness, he was able to accomplish more on the cost side than on the effectiveness side. Using the ingredients approach and somewhat more extensive documentation than that available to King (because of the later date of his research), Barnett was able to make estimates of costs for each model. He, too, relied extensively on documents that suggested the cost elements instead of collecting data directly from school sites. The preferred method is to base costs on the value of the actual ingredients used in an intervention (Levin & McEwan, 2001). Barnett also found that Success for All showed the highest cost per student and Accelerated Schools showed the lowest costs, with the School Development Program occupying the middle position. Barnett made an attempt to determine the effectiveness of the interventions, but found that the available data were insufficient to make direct comparisons. He concluded that all three models show promise of being effective, but was unable to draw more precise conclusions.

Both the King and Barnett studies are pioneering in their

attempts to cope with the tremendous complexities of doing cost-effectiveness comparisons among whole school change projects. Nevertheless, their studies fall short of providing comprehensive cost-effectiveness results because of the inadequacy of existing data and other obstacles that will be developed in the next sections.

Consideration of Resource Reallocation

In some cases a whole school reform will require considerable additional resources beyond those initially deployed to meet its goals. Specific personnel, facilities, equipment, or additional personnel time will prove necessary. But instead of financing these requirements with additional funds, existing resources are reallocated from other uses. For example, a school may be expected to give up other programs and activities to finance the ingredients needed for the reform. To the naive observer the reform may appear to be costless. This assertion assumes that the resources had no productive use whatsoever before they were reallocated. That is, they acquired value only after being redirected to the reform effort.

In order for reallocation to be legitimately costless, it must be shown that there is no loss of other valued outcomes; that is, that the reallocated resources represent a deadweight loss in their previous use, producing nothing of value. But if the resource had any value at all previously, the value should be allocated as a cost to the reform effort. Without a complete mapping and measurement of all valued outputs in assessing effectiveness, it would be impossible to ascertain that cost. When something of value is sacrificed, there is an economic cost. Accordingly, the true cost of the school reform is not only any additional resources that are added, but also the value of all resources that are reallocated from unmeasured outputs to the measured ones. Only if an evaluator can show that none of the other outputs are affected by the reallocation can it be considered costless. More likely the resources were at least somewhat productive in their initial use. As a practical matter, taking resources from one activity (e.g., music or social studies or special education) may solve the challenge of financing new programs. But to assert that there is no cost is incorrect.

Consider a shift in teachers and teacher time from one subject to another. The cost is the value of what is given up in productivity

in the other subject. If time devoted to reading is increased by one hour at the expense of one hour of mathematics, it is likely that the mathematics performance of students will be affected. Thus, the redeployment of resources in this case is not costless, and to make-up any loss in mathematics achievement would likely require replacement of the reallocated resources.[12] Even such "popular" redeployments as gaining resources through reducing extracurricular activities will not be costless if they lead to less student engagement, higher dropout rates, and more student delinquency in the hours that would have been taken up by these activities.

This means that a cost-effectiveness analysis of school reforms that focuses on only a single or a limited set of outcomes by reallocating resources must undertake one of two tasks. First, it can choose to specify the major outcomes of the school, whether they are addressed by the reform or not, and measure the changes in all outputs resulting from the intervention, to ascertain the impact on both the outputs of focus and those from which resources have been reallocated.[13] A more modest approach would be limited to measuring changes only in those outputs that would appear to be impacted by the reallocation to assess how they are affected along with those to which the recources are addressed. This more limited undertaking would require identifying the source of the reallocations and including those outputs in the analysis. Simply assuming, without direct verification, that reallocation is costless and that nothing is given up in that process is inappropriate and almost certainly incorrect.

If the data for making comparisons in effectiveness of whole school reform models are wanting, data on costs are almost nonexistent. There are virtually no systematic studies of costs using the ingredients or resource method other than the attempts by King (1994) and Barnett (1996). Typically, the costs used are limited to those paid to the sponsor of the model for adoption and technical assistance. For example, the New American Schools "price list" for 2001-02 for the school reform models that it represents varies from about $45,000 a year to about $100,000 for schools with about 500 students.[14] This would appear to be about $90 a year to $200 a year per student. The comparison of models by Herman (1999) reports values in this range, but much lower if "current staff are reassigned."

For almost all the models the costs for contracting with the developer are the part of the iceberg that is visible. Below the surface

there are costs of additional personnel to implement the models, such as coaches, teachers, coordinators, and, perhaps, materials, and travel. For example, the Accelerated Schools Project requires a quarter-time external coach and internal facilitator for each school, and most of the other models require a full-time coordinator. Success for All requires teachers who will serve as tutors for at least 30 percent of first graders with 8-11 students per teacher, a minimum of three full-time, additional teachers, and a full-time coordinator of family support. These five positions with salary and fringe benefits comprise an additional cost of $250-300 thousand dollars. Odden, Archibald, and Tysen (2000) suggest that another model, Modern Red Schoolhouse, is even more costly, although the details of the methodology and their application in particular settings (as opposed to theoretical costs) are not identified.

Cost Recovery

Many of the developers of reform models argue that all of these costs can be covered by Federal grants such as Title I or through reallocations. This is a view that is reinforced by a noted expert in financing education, Alan Odden (Odden & Archibald, 2001; Odden et al., 2000). But virtually no attempt is made to consider what is being sacrificed when resources are shifted from one use to another. The result is that these reallocations are treated as costless. In fact, however, many schools find out later that the total cost of reform is much higher in terms of resource requirements than the cost information provided by the reform sponsor if they do not wish to sacrifice other programs.[15]

As discussed above, the cost of an intervention must be based upon the value of all of the resources that are required to replicate it. Once the costs are determined, it will be necessary to figure out how to fund the intervention. Reallocation of resources is a method of financing, but it is not cost free, as some advocates have claimed. The appropriation of all Title I funds for a school reform program devoted to one or two subjects entails sacrifices of other beneficial uses of those funds. These sacrifices may include the loss of benefits from screening and intervention programs for students health needs, psychological services, after-school programs, and academic activities. Cost analysis must take this into account.

Conclusions and Recommendations

The bottom line is that available data are not sufficient to make cost-effectiveness comparisons among the different whole school reform models. The effectiveness results are not based upon a standard sampling strategy among schools. Measures of outcomes favor some models over others and are particularly deficient where reforms are focusing on more than one or two outcomes. Evaluation models differ substantially in both their design and the quality of their implementation. Further, cost data are not based upon a rigorous methodology for identifying the actual resources required for replication and obtaining accurate estimates of their costs. In particular, the recommendations of model sponsors to reallocate resources tend to hide the true cost of the reforms because they ignore what is being given up when resources are reallocated.

This raises the question of what needs to be done to obtain comparability for cost effectiveness purposes.

(1) Combine for comparison those groups of reforms with similar goals rather than impose the same outcome criterion on all models. If reforms aim to increase student discourse, problem solving, research, and artistic endeavors, limiting outcome measures to standardized tests of reading and mathematics will be inappropriate.

(2) Include in the sample a population representative of all attempts to replicate the model or some other consistent criterion, rather than permit evaluators (especially first-party evaluators or reform sponsors) to select their samples on the basis of convenience or ostensible success.

(3) Use a similar set of methodologies in making comparisons, not only in evaluation design, but in details such as how comparison schools are selected in the event of quasi-experimental designs using comparison schools.

(4) Employ a rigorous and systematic methodological approach to cost estimation based upon the state-of-the-art.

(5) Employ third-party evaluators who are both disinterested in the

outcomes and funded by organizations other than an individual reform sponsor.

(6) Increase funding for evaluations of whole school reforms by government agencies and charitable foundations that are ostensibly neutral and do not favor a particular reform for ideological reasons.

Above all, the audiences for such evaluations should be made aware of the flaws in existing comparisons and "marketing" claims of superiority of one model over another. Although the challenges of making truly comparable cost-effectiveness comparisons are great, it is clear that the knowledge base can be improved considerably. At the present time, those who use evaluations of whole school reforms for making adoption decisions should augment the evaluations with other data. These data might include school visits for interviews and observations, as well as broader measures of school success and detailed resource requirements from those sites rather than limiting their perspectives to existing evaluation reports.

ENDNOTES

1. See Northwest Regional Educational Laboratory (2001) for a more nearly complete list with details of each model. See also Desimone (2000) for in-depth descriptions of 24 reform models.

2. Slavin and Fashola (1998) and Herman (1999), the authors of the effectiveness comparisons of the different models, have been associated with a particular reform, Success for All and Roots and Wings, although Herman's study was done subsequently when she was an employee of the American Institutes of Research. It is also important to note that there are far more evaluations of Success for All than of the other reforms because of Robert Slavin's concern with demonstrating the effectiveness of his reform model. In this respect we owe Slavin an important debt. The down-side of that largesse is that the evaluations of Success for All provide many of the examples of issues that are raised in this paper.

3. Slavin and Madden (1999) erroneously refer to evaluations of Success for All as experiments even though they do not use random assignment as the basis for comparison.

4. Even random assignment among "bought-in" schools might not provide appropriate comparisons if the schools that are randomly rejected for school reform experience a "let-down" because of the rejection. See Fetterman (1982).

5. Although all of the models seem to set similar criteria in their printed materials, the verification of buy-in varies profoundly from model-to-model and, perhaps, even school-to-school within models. See Datnow (2000).

6. Olejnik and Algina (2000) provide a cogent presentation on the limits of using measures of effect sizes for comparative studies. They conclude that "...measures of effect size are affected by the research design used" (p. 280). Their cautions on the variability of effect size that is based upon differences in evaluation methods rather than differences in "true" effects are much more extensive than the measurement issue raised here. Also see Lipsey (1999) and Hunter and Schmidt (1994).

7. There has been considerable controversy over what is a third party evaluation, particularly for Success for All/Roots and Wings as reflected in the exchange between a critic, Pogrow (2000), and the founders of Success for All, Slavin and Madden (2000). Slavin and Madden assert that anyone not situat-

ed in their foundation or at their center at Johns Hopkins University are third parties even if they are or have been closely associated with his organizations as collaborators, consultants, or former employees. Most of the Success for All evaluations have been done by persons who have been associated with the organizations of the founder in one or more significant respects.

8. For example, Cooper, Charlton, Valentine, & Muhlenbruck (2000) found that in a meta-evaluation of summer school effectiveness, the internal evaluations produced effect sizes about twice as large as external evaluations.

9. The principal coauthor and two of the others among the five coauthors were affiliated with Success for All/Roots and Wings, which was found to have the largest average effect size across subjects. Slavin and Madden (2000) have argued that the inclusion of two outsiders make this a third-party study.

10. An important point made by Steve Barnett to me is that projects may utilize larger amounts of resources that are donated by other entities, such as contributed time of volunteers or a facility that is provided at no cost to the project, than would be used if they were paid for at full cost. Thus, it is important to ascertain what the necessary ingredients would be in the absence of this type of "distortion," that is, when full costs are taken into account.

11. For example, some of the reform models have persistent high turnover among teachers. This imposes a cost on schools and on districts for teacher selection, training, and personnel accounting that is not captured by typical cost studies. These costs are especially high at a time of teacher shortages.

12. A concrete example is provided in a study of a major school reform that focused on reading. Both instructional time and personnel were redeployed from other subjects and activities to reading. Although reading gains were higher in the intervention school, mathematics gains were higher in the matched comparison school (Jones, Gottfredson, & Gottfredson, 1997).

13. This matches the problem on the output side of valuing total output when it is divided among many non-commensurate dimensions. See Levin and McEwan (2001), Chapter 8.

14. Success for All is a reading program. It charges about $75,000-85,000 a year in the initial year and twice that if Math Wings (a math program) is

added as part of Roots and Wings, or about $150,000 a year. This is about $300 a year per student for contracting with the developer, but not including any costs for additional resources at the school site.

15. In the course of writing this paper I came across an attempt to do a cost-effectiveness analysis of the persisting effects of Success for All relative to reductions in class size and a preschool program (Borman & Hewes, 2001). Although I admire this as a first effort, it suffers from the flaw of both "cost-less" reallocation and of not considering the costs of resources provided by other agencies such as social workers. It is also based upon the initial replications of the project in Baltimore, the site of the sponsoring institution. The substantial assistance by professional and other staff in implementing the project at those sites is not included in the measured costs in this study. In addition, it is inconsistent in comparing its figures with the putative full cost of class size reduction and preschool programs, since each of these could also "reduce" costs through reallocation of resources or through obtaining resources from external sources. Perhaps the largest bias is introduced by charging the full cost of class size reduction to reading, rather than recognizing that class size reductions also improve mathematics (as in the Tennessee experiment) and other subjects and activities. The preschool program also is devoted to a wide range of child outcomes. Previous studies have charged one-third of the cost of class size reduction to the improvement of reading (Levin, Glass, & Meister, 1987). When this adjustment is made, the cost effectiveness estimate for class size reduction is superior to that of Success For All by about 2.5 to 1. For examples of class-size reduction through redeployment of teachers (in one case in a Success for All school), see Miles and Darling-Hammond (1998).

References

Barnett, W.S. (1996). Economics of school reform: Three promising models. In H.F. Ladd (Ed.), *Holding schools accountable* (pp. 299-326). Washington, DC: The Brookings Institution. (ED 396 428)

Begg, C. B. (1994). Publication bias. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399-410). New York: Russell Sage Foundation.

Berends, M., Kirby, S.N., Naftel, S., & McKelvey, C. (2001). *Implementation and performance in New American Schools.* Santa Monica, CA: RAND. (ED 451 204)

Bloom, H., Ham, S., Kagehiro, S., Melton, L., O'Brien, J., Rock, J., & Doolittle, F. (2001). *Evaluating the Accelerated Schools program: A look at its early implementation and impact on student achievement in eight schools.* New York: Manpower Development Research Corporation.

Borman, G.D., & Hewes, G.M. (2001). *The long-term effects and cost-effectiveness of Success for All.* Available: www.successforall.net/resource/researchpub.htm

Boruch, R. (1997). *Randomized experiments for planning and evaluation: A practical guide.* Thousand Oaks, CA: Sage.

Bowles, S., & Gintis, H. (2000). Does schooling raise earnings by making people smarter? In K. Arrow, S. Bowles, & S. Durlauf (Eds.), *Meritocracy and economic inequality* (pp. 118-36). Princeton: Princeton University Press.

Calaway, F. (2001). *Evaluation of the comprehensive school reform models in the Memphis City Schools.* Memphis: Memphis City Schools, Office of Research & Evaluation.

Comer, J. (1988, November). Educating poor minority children. *Scientific American, 259*(5), 42-48. (EJ 386 132)

Cook, T.D., & Campbell, D.T. (1979a). *Quasi-experimentation: Design & analysis for field studies.* Chicago: Rand McNally.

Cook, T.D., & Campbell, D.T. (Eds.). (1979b). *Quasi-experimentation: Design & analysis issues for field settings.* Boston: Houghton Mifflin.

Cook, T.D., Farah-Naaz, H., Phillips, M., Settersten, R.A.; Shagle, S.C., & Degirmencioglu, S.M. (1999, Fall). Comer's School Development Program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal, 36*(3), 543-97. (EJ 613 947)

Cook, T.D., Murphy, R.F., & Hunt, H.D. (2000, Summer). Comer's School Development Program in Chicago: A theory-based evaluation. *American Educational Research Journal, 37*(2), 535-97. (EJ 624 133)

Cooper, H., Charlton, K., Valentine, J.C. & Muhlenbruck, L. (2000). Making the most of summer school: A meta-analytic and narrative. *Monographs of the Society for Research in Child Development, 65*(1), 1-118. (EJ 630 022)

Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms, 1890-1990* (2nd Ed.). New York: Teachers College Press. (ED 388 482)

Datnow, A. (2000, Winter). Power and politics in the adoption of school reform models. *Educational Evaluation and Policy Analysis, 22*(4), 357-74.

Desimone, L. (2000). *Making comprehensive urban school reform work.* Urban Diversity Series No. 112. New York: Teachers College, ERIC Clearinghouse on Urban Education and the Institute for Urban and Minority Education. (ED 441 915)

Doolittle, F. (2001, April). *Using interrupted time-series analysis to measure the impacts of Accelerated Schools on the performance of elementary school students.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle. New York: Manpower Development Research Corporation.

Edmonds, R. (1979, October). Effective schools for the urban poor. *Educational Leadership, 37*(1), 15-18, 20-24. (EJ 208 051)

Fetterman, D.M. (1982). Ibsen's baths: Reactivity and insensitivity. *Educational Evaluation and Policy Analysis, 4*(3), 261-79.

Finnan, C., & Levin. H. (2000). Changing school cultures. In H. Altrichter & J. Elliott (Eds.), *Images of educational change* (pp. 87-99). Philadelphia: Open University Press.

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

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Grissmer, D. (2002). Cost-effectiveness and cost-benefit analysis: The effect of targeting interventions. In H. Levin & P. McEwan (Eds.), *Cost-effectiveness and educational policy.* Yearbook of the American Education Finance Association. Larchmont, NY: Eye on Education.

Hargreaves, A., Lieberman, A., Fullan, M., & Hopkins, D. (Eds.). (2000). *International Handbook of Educational Change* (Parts One and Two). Boston: Kluwer.

Herman, R. (1999). *An educators' guide to schoolwide reform.* Arlington, VA: Educational Research Service.

Hopfenberg, W.S., Levin, H., Chase, C., Christensen, S.G., Moore, M., Soler, P., Brunner, I., Keller, B., & Rodriguez, G. (1993). *The Accelerated Schools resource guide.* San Francisco: Jossey-Bass. (ED 365 758)

Hunter, J. E., & Schmidt, F.L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis.* New York: Russell Sage Foundation.

Inkeles, A. (1966).The socialization of competence. *Harvard Educational Review, 36*(3), 265-83.

Jones, E., Gottfredson, G., & Gottfredson, D. (1997). Success for some: An evaluation of the Success for All program. *Evaluation Review, 21*(6), 599-607.

Kane, T.J., & Staiger, D.O. (2001). *Improving school accountability measures.* Working Paper 8156. Cambridge, MA: National Bureau of Economic Research. Available: www.nber.org/papers/w8156

Kerbow, D. (1996). *Patterns of urban student mobility and local school reform.* Technical Report No. 5. Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk. (ED 402 386)

King, J.A. (1994, Spring). Meeting the educational needs of at-risk students: A cost analysis of three models. *Educational Evaluation and Policy Analysis, 16*(1), 1-19.

Levin, H. (1991). Cost-effectiveness at quarter century. In M.W. McLaughlin & D.C. Phillips (Eds.), *Evaluation and education: At quarter century.* Ninetieth Yearbook of the National Society for the Study of Education (Part II, pp. 189-209). Chicago: University of Chicago Press.

Levin, H. (1997, June). Raising school productivity: An x-efficiency approach. *Economics of Education Review, 16*(3), 303-11. (EJ 547 333)

Levin, H. (2001). Waiting for Godot: Cost-effectiveness analysis in education. In R.J. Light (Ed.), *Evaluation findings that surprise.* New Directions for Evaluation, 90 (pp. 55-68). San Francisco: Jossey-Bass.

Levin, H.M., Glass, G., & Meister, G. (1987, February). Cost-effectiveness of computer-assisted instruction. *Evaluation Review, 11*(1), 50-72. (EJ 353 322)

Levin, H.M., & McEwan, P.J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd Ed.). Thousand Oaks, CA: Sage.

Lipsey, M.W. (1999). Can rehabilitative programs reduce the recidivism of juvenile offenders? *The Virginia Journal of Social Policy and the Law, 6*(3), 611-41.

Mac Iver, M., Stringfield, S., & McHugh, B. (2000). *Core Knowledge curriculum: Five-year analysis of implementation and effects in five Maryland schools.* Report No. 50. Baltimore: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk.

McCain, L.J., & McCleary, R. (1979). The statistical analysis of the simple interrupted time-series quasi-experiment. In T.D. Cook & D.T. Campbell (Eds.), *Quasi-experimentation: Design & analysis issues for field settings* (pp. 233-93). Boston: Houghton Mifflin.

McGaw, B. (1994). Meta-analysis. In T. Husen & T.N. Postlethwaite (Eds.), *The international encyclopedia of education* (Vol. 7, 2nd Ed., pp. 3775-84). New York: Pergamon.

McNeil L., & Valenzuela, A. (2000). *The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric.* Unpublished manuscript. (ED 443 872)

Miles, K.H., & Darling-Hammond. L. (1998, Spring). Rethinking the allocation of teaching resources: Some lessons from high performing schools. *Educational Evaluation and Policy Analysis, 20*(1), 9-29.

Millsap, M. A., Chase, A., Obeidallah, D., Perez-Smith, A., Brigham, N., & Johnston, K. (2000). *Evaluation of Detroit's Comer Schools and Families Initiative.* Final Report. Cambridge, MA: Abt Associates.

Northwest Regional Educational Laboratory. (2001). *Catalog of school reform models.* Portland, OR: Author. Available: www.nwrel.org/scpd/catalog

Odden, A., & Archibald, S. (2000). *Reallocating resources: How to boost student achievement without asking for more.* Thousand Oaks, CA: Corwin. (ED 450 441)

Odden, A., Archibald, S., & Tychsen, A. (2000, Winter). Can Wisconsin schools afford comprehensive school reform? *Journal of Education Finance, 25*(3), 323-42.

Olejnik, S., & Algina, J. (2000, July). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*(3), 241-86.

Pogrow, S. (2000, September). Success for All does not produce success for students. *Phi Delta Kappan, 82*(1), 67-81. (EJ 612 901)

Ross, S., & Smith, L. (1994, November). Effects of the Success for All model on kindergarten through second-grade reading achievement, teachers' adjustment, and classroom-school climate at an inner-city school. *Elementary School Journal, 95,* 121-38. (EJ 493 622)

Ross, S. M., Wang, L W., Sanders, W.L., Wright, S.P., & Stringfield, S. (1999). *Two- and three-year achievement results on the Tennessee Value-Added Assessment System for restructuring schools in Memphis.* Memphis: University of Memphis, Center for Research in Educational Policy.

Rumberger, R.W. (2001). *Mobility and student outcomes.* Arlington, VA: Education Research Service.

Sanders, W.L., & Horn, S.P. (1995). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. In A.J. Shinkfield & D.L. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp. 337-50). Boston: Kluwer. (ED 435 632)

Sarason, S.B. (1982). *The culture of the school and the problem of change* (2nd Ed.). Boston: Allyn & Bacon.

Scriven, M. (1976). Evaluation bias and its control. In G.V. Glass (Ed.), *Evaluation studies review annual* (Vol. 1, pp. 119-39). Beverly Hills, CA: Sage.

Slavin, R.E., & O.S. Fashola. (1998). *Show me the evidence! Proven and promising programs for America's schools.*Thousand Oaks, CA: Corwin. (ED 421 488)

Slavin, R.E., & Madden, N.A. (1999). *SUCCESS FOR ALL/ROOTS & WINGS: Summary of Research on Achievement Outcomes.* Report No. 41.Baltimore: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk. (ED 438 363)

Slavin, R.E., & Madden, N.A. (2000, September). Research on achievement outcomes of Succcess for All: A summary and response to critics. *Phi Delta Kappan, 82*(1), 38-40, 59-66. (EJ 612 899)

Slavin, R.E., Madden, N.A., Karweit, N.L., Livermon, B.J., & Dolan, L. (1990, Summer). Success for All: First-year outcomes of a comprehensive plan for reforming urban education. *American Educational Research Journal, 27*(2), 255-78. (EJ 414 291)

Stringfield, S., Datnow, A. Borman, G., & Rachuba, L. (1998). *National evaluation of Core Knowledge sequence implementation: Final report.* Baltimore: Johns Hopkins University, Center for the Social Organization of Schools. (ED 451 282)

Stringfield, S., Millsap, M.A., & Herman, R. (1998). Using "promising programs" to improve educational processes and student outcomes. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins, *International handbook of educational change* (Part Two, pp. 1314-38). Boston: Kluwer.

Suchman, E. (1971). Evaluating educational programs. In F. Caro (Ed.), *Readings in evaluation research* (pp. 43-48). New York: Russell Sage Foundation.

Summers, A., & Wolfe, B. (1977, September). Do schools make a difference? *American Economic Review, 67*(4), pp. 639-52.

Supovitz, J. A., Poglinco, S. M., & Snyder, B.A. (2001). *Moving mountains: Successes and challenges of the America's Choice comprehensive school reform design.* Philadelphia: University of Pennsylvania, Center for Policy Research in Education.

Wilson, J. (1973). On Pettigrew and armor: Afterword. *The Public Interest, 30,* pp. 132-34.

Biography of Author

Henry M. Levin is the William Heard Kilpatrick Professor of Economics and Education and Director of the National Center for the Study of Privatization at Teachers College, Columbia University. He is also the David Jacks Professor of Higher Education and Economics, Emeritus, at Stanford University. Dr. Levin is the Founding Director of the Accelerated Schools Project, a national school reform that was established in 1986. He is a specialist in the Economics of Education and School Reform and is the author or editor of 18 books and about 300 articles. His latest books, co-authored with Patrick McEwan, are *Cost Effectiveness Analysis: Methods and Applications* (Sage Publications, 2001) and *Cost Effectiveness and Educational Policy* (Eye on Education, 2002).

ERIC CLEARINGHOUSE ON URBAN EDUCATION
INSTITUTE FOR URBAN AND MINORITY EDUCATION
TEACHERS COLLEGE, COLUMBIA UNIVERSITY
NEW YORK, NY 10027

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").