

Session 2 Discussion and Synthesis Comments for Symposium on the Use of School-
Level Data for Evaluating Federal Education Programs^{*}

Laura Hamilton, RAND Corporation

December 8, 2005

Thank you for the opportunity to participate in this symposium. I enjoyed the papers and I think the discussion so far has been very productive. I'll start by saying a few words about each of the papers, and then I'll discuss some themes that cut across all or most of them. I'll focus my comments on issues related to measurement and program evaluation and will try not to overlap with the topics that will be covered by Bob Linn and Judy Singer.

Elizabeth Stuart

Elizabeth Stuart's paper provides a clear, cogent discussion of causal inference. The language is largely accessible to non-technical audiences; this piece represents the kind of discussion that could be provided to educators and administrators who are increasingly being expected to make decisions based on the results of research. In particular, it may help discourage an assumption that any study that does not use a randomized design is necessarily invalid for estimating causal effects. Stuart nicely

^{*} Copyright © 2005 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.

Opinions and statements included in the paper are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Planning Committee for the Symposium on the Use of School-Level Data in Evaluating Federal Education Programs or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.

describes the conditions necessary for making causal inferences and the degree to which various designs approximate these.

Her paper calls attention to the need to define what that control condition is, a topic to which I return later. Her simple reading curriculum example illustrates that attempting to interpret an effect size in the absence of detailed information about the study is risky, and that meta-analyses need to carefully consider specific study features before combining results to produce a single estimate.

Her discussion of the stable unit treatment value assumption (SUTVA) is important for evaluators to keep in mind. Violations of SUTVA are common in education, as Stuart notes. Stuart describes spillover effects that can occur as a result of interaction between treatment and control group members, but it should be noted that violations can occur even in the absence of direct interaction between treatment and control. Consider the competitive effects that are often discussed in the choice literature. Districts or schools may adopt certain practices that are part of the treatment without necessarily having any direct contact with treatment sites. The assumption is also violated by frequent variation in the nature of the treatment and control conditions.

The paper provides a nice discussion of propensity score matching. In the context of the SSASD, I wonder whether we typically have sufficient covariates to do an adequate match. The prospect of linking to other information such as census is helpful. SASS does not include information on all schools but does provide some potentially important covariates such as information about school climate and crime.

Yeow Meng Thum

The idea of gross productivity indicators is intriguing, and might provide an effective antidote to all-too-common tendencies to make causal conclusions based on growth data. My primary concern is with interpretation. The proposed index is designed to prevent overly strong attributions to teachers or schools, but it is almost certainly going to be difficult to communicate what this is in a way that will prevent confusion and misuse. The idea of gross productivity doesn't fit with most users' notions of what "accountability" means.

Thum notes the importance of having a valid interval scale and tests that are well-equated across grades (though "linked" might be a better term than "equated"). Although many current testing programs use vertical or developmental scales that appear to meet these conditions, there are concerns with the extent to which these scales provide valid information about growth on a consistent construct. Researchers such as Joseph Martineau and Bill Schmidt have shown how the nature of what is measured changes across grades and how these changes can affect estimates of teacher or school performance. This is particularly problematic for grade levels that are far apart or for curricula that are not cumulative (e.g., most science curricula). In addition, changes in scores do not mean the same thing at all points in the achievement distribution, which underscores the importance of matching on student achievement.

Another problem with vertical scales is that variance trends differ depending on scaling approach. There has been intense debate in the psychometric community about this problem, which can affect conclusions based on growth models. A third issue is that it is not equally difficult to achieve a certain amount of movement at all points in the

distribution. For example, improving the reading performance of a 6th grader reading at a 3rd grade level is more difficult than doing so for a 6th grader reading at the 6th grade level. All of these problems threaten the validity of inferences about growth across a population and for individual units.

One of the drawbacks of the kinds of individual growth models that Thum describes is that there is typically a large amount of missing data, some of it as a result of testing schedules. In many states, for example, elementary schools do not administer a uniform statewide test until the end of grade 3. If growth for the school is measured using spring of grade 3 as a starting point, a large proportion of students' time in that school will be excluded from the estimate (e.g., in a typical K-5 school, a measure of change from spring of grade 3 to spring of grade 5 will include only 1/3 of the time the student spent there). If a school increases its effectiveness at improving learning in the early grades, conclusions about growth in the school or about changes in growth rates over time will be distorted.

Finally, conclusions about improvement in performance of successive cohorts might depend on factors unrelated to improvements in achievement, such as increasing test familiarity (which might lead to score inflation) or changes in test format.

Michael Scriven

This paper provides a nice discussion of how causal inferences are made in other scientific fields; random experiments are not the only basis. Scriven also points out some problems that are inherent in RCTs in education, such as the lack of a double blind design.

The discussion of elimination analysis illustrates the complexity involved in making inferences about effectiveness. For this process to work there need to be well-understood and tested theories about relationships between treatments and outcomes, and highly detailed outcome data. The process also requires strong diagnostic ability on the part of the person interpreting the results. These conditions are unlikely to be met currently, but perhaps could be met with appropriate investment in resources and human capital. In particular, there have been developments in the technology of assessment that might provide ways to create better diagnostic information than is typically provided by state assessments.

Scriven makes some important points about the value of case study research. One drawback to this type of research is that it is often difficult to get a representative sample of schools to participate, because schools' willingness to let researchers in is likely to be related to implementation, achievement, or other school characteristics. Even though case study information is likely to be extremely useful for understanding the nature of the intervention and the mechanisms through which it might contribute to student achievement, caution is warranted when attempting to generalize beyond the types of schools or classrooms included in the case study sample.

Gary Miron

This paper provides a useful framework for classifying study designs. Miron notes threats to validity of inferences based on successive cohorts. The most important one is probably changes in student populations, and Miron discusses a few ways to address these. His discussion echoes some of the other authors' suggestions to create a

comparison group that is similar to treatment group. Some of the suggested approaches do not address unmeasured background characteristics (e.g., student or family motivation) and may depend on assumptions that are difficult to confirm (e.g., linearity), but they provide a way to create approximately equivalent groups.

There is one issue that is particularly relevant to pre-post designs. When measuring gains, it is difficult to estimate a treatment effect in the absence of a true baseline. For example, RAND’s evaluation of Edison schools involved examining achievement changes starting with a pre-Edison baseline, but those scores were not available for new start-up schools. For those schools it was necessary to use achievement test results from the spring of the first year of Edison management. The problem with the first-year baseline is that it results in a loss of information about what occurred during the first year, and this can distort results. We saw declines in achievement during Edison’s first year in conversion schools but had no way of verifying whether these declines also characterized the performance of start-up schools. The problem of missing baselines is inherent in efforts to examine the performance of start-up schools (such as charters) using school-level data.

Miron describes an interesting application of odds ratio analysis. Like all analyses that focus on percent proficient, it is important to address distributional assumptions. Also consider behavioral responses arising from the “bubble kids” phenomenon, which I discuss later.

Miron’s suggestions for synthesizing data are useful, though determining the optimal way to weight various types of studies and to take into account idiosyncratic

features of studies that might make them more or less valid than the typical study in their category is challenging.

Overarching themes, or considerations when conducting and interpreting analyses of program effects using school-level data

Now I'll discuss a few general themes that cut across all or most of the papers. These comments reflect a combination of lessons I took away from the papers and my own experience with using test-score data.

1. Importance of understanding both treatment and control conditions

When estimating the effects of any type of treatment using a comparison with an untreated group, it is essential to understand the characteristics of the units assigned to each of those groups. Consider the charter school example. Charter schools vary on a number of dimensions including students served, level of autonomy, and funding.

Enrollment in a charter school does not provide exposure to any specific treatment; charters use a wide variety of curricula, approaches to scheduling, etc. This variation makes it difficult to satisfy many of the assumptions that underlie causal modeling (such as SUTVA), and hinders efforts to interpret the results.

This type of variation is not only inherent in interventions such as charter schools, which by design often encompass a great deal of flexibility, but is also likely to characterize even well-defined interventions such as a specific reading curriculum. Schools will vary in the quality of teachers who are using the curriculum, the adequacy of resources to support its implementation (e.g., classroom space), the degree to which it is

being used as its developers intended, etc. It is difficult and expensive to obtain information on these factors but this information is critical if we are to fully understand the nature of a particular program effect or lack of effect.

The same problem applies to the comparison schools, and in some ways is even more serious there because there may be a wider range of actions taking place and less leverage for collecting information on these actions.

A clear definition of what is meant by “comparison” or “control” is important for informing the selection of a comparison group. For example, in RAND’s Edison study we needed to decide whether to select comparison schools from within the same districts as the Edison schools or from outside those districts. If we think schools must be matched on issues related to district context, then within-district matches might be more appropriate. However, if we are worried about competitive effects or selection effects, we might be better off with outside matches. The standard advice to “match on as many characteristics as possible” might not always lead to the most accurate estimates of program effects, particularly when some of the obvious matching variables turn out to be important parts of the treatment.

The problem of finding an appropriate control group might become more serious as schools become better at identifying scientifically based interventions. If all schools eventually identify and adopt effective curricula, then any test of a new curriculum will necessarily compare it against an intervention already proven to be effective. In that case, a finding of “no difference” may be interpreted as evidence of effectiveness.

Finally, both treatment and control schools are affected by outside-of-school influences, including summer effects. These are nearly impossible to address fully with

existing data and it's difficult to imagine a data collection effort that would address them while being reasonably affordable. Some interventions might directly address outside influences, whereas others might indirectly address them. This needs to be understood when interpreting program effects.

2. The importance of understanding users' inferences

The papers presented today addressed the use of test scores for both research and accountability. For both of these uses it's important to understand the inferences that users make on the basis of information they receive. For example, when a parent interprets an accountability index, he or she makes inferences about the generalizability of any gains (e.g., the parent might assume that his child's SAT scores will rise because his school's state test scores increased, or he might limit his inferences to material included in the state standards). Similarly, when interpreting research results, users make inferences about what they mean and what kinds of decisions they can support. For example, what can users conclude from a finding that says charter schools on average are outperforming conventional schools? Different users' inferences and uses are likely to vary—a district superintendent who is considering converting schools to charters will use the results differently from a parent deciding where to send her child.

In fact, it is possible that cross-sectional information from a single point in time might be more useful to some users than change scores, at least for some purposes. Consider the parent who believes in peer effects or who thinks schools will tailor their instruction to the average or minimum ability levels of the students. Of course, whether

this information is perceived as useful and whether there should be policies to support providing it are two different questions.

Decisions about what to control for should depend in part on the inferences that the study or accountability system is intended to support. For example, controlling for demographics is generally appropriate for research, but may or may not be appropriate for an accountability system, depending on its goals. It could be perceived as setting lower standards for some students than for others. The line between research and accountability is not always clear; e.g., a district might want research on the effects of a new reading curriculum but might also use the resulting information to make decisions about individual schools or teachers. Those responsible for communicating study results or accountability information need to address the needs of likely users and work to prevent misinterpretation and misuse.

3. Match between tests and instruction

For test scores to provide valid information about school or teacher performance, the test needs to be sensitive to the curriculum and instruction that are intended to be provided by the program. If teachers are fully implementing the program but the test does not adequately match the instruction they are providing, the results could provide misleading information about program effectiveness.

An example of a problem that can arise from mismatch stems from the nature of mathematics instruction in middle schools, where there is often tracking (e.g., some students might take algebra whereas others take a general math class). Most state math tests for the middle school grades cover a range of topics. Therefore these tests might be

more appropriate for measuring the effects of instruction in the general math class than in the algebra class. Both between- and within-grade variations in curricula can distort estimates of educational effectiveness. High schools in states where specific end-of-course tests are administered present a different problem—the tests might be designed to match the instruction and curriculum, but might not be easily comparable with one another.

Of course, one could take the view that in order to be deemed effective the intervention needs to be shown to improve scores on tests aligned with state standards, regardless of whether those tests match the curriculum, but that view needs to be made clear when presenting the results. In addition, use of state accountability tests raises concerns about score inflation, which I discuss next.

4. Behavioral responses to testing

Differences in behavioral responses to testing can lead to differences in the extent to which scores are inflated across units (schools, classrooms) and across treatment conditions. Score inflation is a fairly well-known phenomenon, but it's often believed that there are simple fixes, such as changing the test form each year. While these steps can help, they do not guarantee that inflation won't occur. Sometimes there are more subtle changes that can be made, such as inspecting tests to ensure that a particular topic isn't always asked about in the same way or that the test doesn't use particular item styles or formats excessively. In addition it's important to understand factors that might lead to differences in teachers' propensity to inflate scores, such as different levels of incentives in treatment and control groups.

One specific type of behavioral response is related to the use of a percent-proficient-or-above metric for reporting and evaluating school performance. When accountability systems rely on this metric, it's not uncommon for teachers to focus on students with the highest perceived probability of moving above the cut score. These students are often referred to as “bubble kids,” and surveys have shown that this type of reallocation of resources and attention is a common response to state accountability systems. Changes in percent proficient might therefore be inflated relative to changes throughout the score distribution.

A related problem is the extent to which differences in students' motivation can adversely affect the validity of inferences based on test scores. Students might not be motivated to do their best work when taking state accountability tests, and the degree of motivation might vary as a function of state or district policy (e.g., whether test scores affect decisions about promotion to the next grade) or other factors that are more difficult to measure. Some understanding of the likely effects of student motivation in both treatment and control schools is important for interpreting program effects.

5. Use of a percent above cut (PAC) metric

Several authors noted limitations associated with using a PAC (in addition to the bubble kid problem described above). In particular, it masks growth when changes do not cross over the cut score, and may exaggerate growth in cases where the cut score is crossed. There are other problems, including variation across states in where the cut score is set, which can lead to strange distributional problems. Movement from below to above PAC means something different across states and even across grade levels within a

testing program. For example, it might be more difficult to achieve a 10% increase in percent proficient in one testing context than in another. PAC measures can also distort information about differences in achievement across groups; this problem is especially relevant today given the emphasis on reducing gaps in performance among racial/ethnic and socioeconomic groups.

Suggestions for Improved Data Collection

There are several steps that can be taken in the data collection process to address some of the problems that the authors, panelists, and other symposium participants have raised. Below I discuss a few.

Ideally, we would gather data on individual students, linked over time and to their teachers. Analyses using individual student data would go a long way toward addressing many of the problems that have been discussed in this symposium. This type of data is becoming increasingly available as states improve their data systems and experiment with growth models, and several states have received grants and other forms of assistance to assemble individual-level databases. An effort similar to the SSASD but focused on assembling states' student-level data as they become available would have an enormous payoff.

Obtaining links to teachers is more difficult, for both political and technical reasons. For example, even when schools maintain such links, they often provide misleading information, such as when team teaching occurs or when students in elementary schools have different teachers for different subjects. Obtaining these links, when possible, is one way to measure within-school and between-school variation in

implementation and context. Efforts to link student and teacher information could take advantage of the work that's currently being conducted to improve the measurement of instructional practice.

Regardless of whether data are gathered at the individual student, teacher, or school level, it would be desirable to have as much information as possible on student achievement and on program implementation. Regarding the former, a large number of symposium participants noted the importance of supplementing information on performance levels with scale scores that can be used to measure achievement and growth at all points in the test-score distribution. Performance levels are of some utility, but provide incomplete or even misleading information for some purposes. It would also be useful to gather any available information on performance on subscores of the achievement tests, since research has shown that conclusions about program effects or educator effectiveness can sometimes vary across subscores. This information can be especially important when evaluating the effects of a program that is intended to promote a specific set of skills (e.g., a mathematics intervention that is focused on computational skills).

One of the problems addressed in other sessions during this symposium (e.g., see Bob Linn's paper and David Thissen's response) is the lack of comparability of assessment systems across states. I'll leave the discussion of how to address this to others, but will note that it might be beneficial for the SSASD or other state test-score databases to include information on each state's testing program. Some of this information could be in the form of quantitative indicators, but some of it might be more effectively communicated through a state profile that describes the testing program.

There are a number of items that could be included: descriptions of the methods used for scaling the test and for setting cut scores; descriptions of the item formats used; evidence of alignment between the test and the standards; policies for inclusion and for provision of testing accommodations; and any other information to help users understand the context surrounding testing, such as the likelihood that students are motivated to perform well (e.g., through information on whether test scores are used to make decisions about individual students) and the likelihood that scores will be inflated (e.g., through information on how test forms change each year). These state profiles could help users of the data understand features of each state's testing program that might affect results.

Of course there is probably a nearly unlimited amount of additional information that could contribute to more effective use of state test-score data, but there are significant costs associated with any new data collection. Decisions about what to include will inevitably be influenced by resource constraints. However, some of the suggestions made during the course of this symposium are reasonably inexpensive to implement. Others are costly, but the benefits of some of these might ultimately outweigh the costs.